INÊS CAMPOS MONTEIRO SABINO DOMINGUES

An automatic mammogram system: from screening to diagnosis

Ph.D. Thesis

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

May 2015

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

An automatic mammogram system: from screening to diagnosis

Inês Campos Monteiro Sabino Domingues

Programa Doutoral em Engenharia Eletrotécnica e de Computadores

Supervisor: Jaime dos Santos Cardoso (PhD)

Co-supervisor: Maria João Viseu Botelho Cardoso (MD, PhD)

May 2015

© Inês Domingues, 2015

Resumo

O cancro mais frequente nas mulheres de todo o mundo é o cancro da mama, sendo a principal causa de morte por cancro em mulheres, sobretudo na faixa etária dos 40 aos 55 anos de idade. A mamografia de rastreio é realizada na população assintomática para detetar sinais precoces de cancro da mama, tais como massas, calcificações, assimetria e distorção arquitetural. A mamografia de diagnóstico é realizada em pacientes que já tenham demonstrado achados clínicos anormais. Tanto as imagens de rastreio como as de diagnóstico são interpretadas por radiologistas que inspecionam visualmente as mamografias. A inspeção manual é uma tarefa cansativa e propensa a erro humano. Desta forma, o desenvolvimento de técnicas automáticas de deteção e diagnóstico (*Computer-Aided Detection and Diagnosis* - CAD) tem sido incentivado. A presente tese descreve um conjunto de contribuições para o desenvolvimento de métodos de processamento de imagem e de aprendizagem automática cujo objetivo é auxiliar radiologistas na análise de imagens de mamografia.

Foram feitas contribuições nas diversas fases de processamento: (1) pré-tratamento, (2) rastreio, (3) deteção de zonas suspeitas, (4) caracterização das regiões suspeitas, e (5) classificação. Todas as técnicas foram exaustivamente avaliadas utilizando uma base de dados de imagens digitais de mamografia, que, juntamente com as imagens, contém meta-informação como a densidade da mama, avaliação BI-RADS (*Breast Imaging Reporting And Data System*) e segmentações muito precisas das regiões suspeitas. Esta base de dados é denominada INbreast.

A manipulação prévia da imagem pode ter um forte impacto no sucesso das tarefas subsequentes. Um pré-processamento típico aplicado a imagens de mamografia é a remoção do músculo peitoral. São apresentados nesta tese dois métodos para a segmentação do músculo peitoral, denominados caminho mais curto em coordenadas polares (SPPC) e caminho mais curto com pontos finais aprendidos por SVMs (SPLE).

Depois da eliminação do músculo peitoral, as imagens passam para a fase de rastreio. Nesta fase, as mamografias são inicialmente classificadas como densas ou não-densas. De seguida, para cada tipo de mama, é desenhado um classificador com o objetivo de determinar se a imagem é suspeita. Um grande conjunto de características, em combinação com diversos tipos de classificadores foram extensivamente avaliados. Para a classificação da densidade, os melhores resultados foram obtidos com o classificador k-vizinhos mais próximos (kNN) e usando um vetor de características de tamanho 6, composto por características estatísticas extraídas de ambas as vistas tipicamente adquiridas. Os melhores classificadores selecionados para a classificação das imagens como suspeitas ou não suspeitas são também kNNs mas usando características de Gabor.

Mulheres com mamografias consideradas como não suspeitas voltam ao esquema habitual de vigilância/rastreio mamográfico recomendado pelo respectivo país. Os exames suspeitos passam para a fase de diagnóstico. Durante esta análise pode ser útil dirigir a atenção do especialista para possíveis regiões problemáticas na imagem. Os dois achados mais comuns observados em imagens de mamografia são calcificações e massas. Devido às suas diferentes características (tamanho, intensidade, forma, contraste na fronteira, etc.) foram utilizados métodos diferentes para a deteção

de cada achado. Um algoritmo baseado em surpresa Bayesiana foi desenvolvido para a deteção de calcificações, enquanto um filtro de Iris seguido por uma técnica de segmentação de contornos fechados foi utilizado para a deteção e segmentação de massas.

Tanto as calcificações como as massas, quando existentes, podem ser benignas ou malignas. A classificação BI-RADS descreve como fatores importantes para a determinação de malignidade a distribuição e morfologia das calcificações e as margens, forma e densidade das massas. Foi realizada uma revisão da literatura e observou-se uma falta de consenso sobre qual o conjunto mais apropriado para a caracterização destes achados. Desta forma, uma grande parte das características descritas na literatura foi avaliada utilizando dados da INbreast e a correlação de Pearson, correlação de distância e o coeficiente de informação máxima. Estas métricas foram usadas para selecionar subconjuntos apropriados de características.

Com base no nível de suspeita, as lesões podem ser classificadas numa das sete categorias BI-RADS: 0 quando o exame não é conclusivo, 1 para nenhum achado, 2 para achados benignos, 3 para achados provavelmente benignos, 4 para resultados suspeitos, 5 quando há uma grande probabilidade de malignidade, e 6 para cancro comprovado. Quando um exame apresenta mais do que um achado, a classificação BI-RADS presente no relatório médico corresponde à classe do achado com o BI-RADS mais alto. Desta forma, as configurações de aprendizagem automática estudadas na literatura não representam toda a informação disponível. Um novo paradigma de aprendizagem, denominado aprendizagem max-ordinal (MOL), é proposto nesta tese com o objectivo de fazer um melhor uso da informação disponível. Este paradigma encontra-se entre a classificação supervisionada e semi-supervisionada. Para cada observação, existe informação sobre o rótulo. No entanto, num subconjunto exemplos, este conhecimento é incompleto e corresponde à classificação do pior caso observado. A formalização do paradigma de aprendizagem max-ordinal levou a dois novos sistemas de aprendizagem, MOL.LA e MOL.CD. O MOL.CD é uma espécie de *coordinate descent* no espaço dos modelos, enquanto que no MOL.LA o foco está no particionamento das instâncias de treino em dois subconjuntos. A avaliação experimental mostrou que as metodologias desenvolvidas apresentam melhores resultados do que o uso de metodologias de aprendizagem típicas.

Todas as técnicas descritas foram cuidadosamente avaliadas tanto de forma independente (supondo que todas as etapas anteriores estão corretas) como em conexão. Concluiu-se que a deteção do músculo peitoral, o rastreio, a segmentação de massas, a extração de características e a classificação BI-RADS estão prontos a ser utilizados na prática. Os algoritmos de deteção de calcificações e massas, no entanto, precisam de ser melhorados de modo a obter sensibilidades mais altas com menos falsos positivos.

O impacto da pesquisa realizada será refletido na capacidade de melhorar a qualidade da deteção de cancro da mama. Assim, o tempo de produção do diagnóstico é menor, melhorando as possibilidades de tratamento da paciente e consequentemente o seu bem-estar físico e psicológico. O radiologista será também beneficiado, podendo concentrar-se nos casos mais complexos.

Abstract

Breast cancer is the most common type of cancer in women worldwide, and the leading cause of death from cancer in women, especially those between 40 and 55 years of age. Screening mammography is performed in the asymptomatic population to detect early signs of breast cancer such as masses, calcifications, bilateral asymmetry and architectural distortion. Diagnostic mammography is performed on patients who have already demonstrated abnormal clinical findings. Both screening and diagnostic mammography are traditionally performed by radiologists who visually inspect mammograms. Manual inspection is a tiring and tedious task prone to human error. In this way, the search for Computer-Aided Detection and Diagnosis (CAD) techniques has been encouraged. The present thesis describes an effort to develop image and machine learning methods to help radiologists in the analysis of mammogram images.

Contributions were made in the different phases, including: (1) pre-processing, (2) screening, (3) detection of suspicious regions, (4) characterization of suspicious regions, and (5) classification. All the techniques were thoroughly evaluated using a database of full field digital mammogram images that, along with the images, contains meta-data information like breast density, BI-RADS (Breast Imaging Reporting And Data System) assessment and very accurate segmentations of suspicious regions. This database is known as INbreast.

Image manipulation can have a strong impact on the success of subsequent tasks. A typical pre-processing applied to mammogram images is the removal of the pectoral muscle region. Two methods for segmentation of the pectoral muscle are presented in this thesis, namely polar coordinates and the shortest path (SPPC) and shortest path with endpoints learnt by SVMs (SPLE).

After pectoral muscle subtraction, the mammogram exam goes through the screening block. Since it has been observed that CAD performance depends on breast density, breasts are first classified as fatty or dense. Then, for each breast type, a specific classification block is designed to determine if the breast exam is suspicious. An extensive evaluation by testing a large set of features in combination with several classifiers was performed. The best density classification results were achieved with a k-Nearest Neighbours (kNN) classifier and using a feature vector consisting of statistic features extracted from both views. The best classifiers selected for the classification of images as suspicious or non-suspicious were also kNNs, but using Gabor features.

Based on the outcome of the screening stage, non-suspicious patients return to the normal screening program advised by different countries while suspicious exams are sent to diagnosis. During this analysis it may be useful to direct the attention of the specialist to regions of the image that may be problematic. The two most common findings seen in mammogram images are calcifications and masses. Due to their different characteristics (size, intensity, shape, border contrast, etc.) different methods were used for each type of finding. An algorithm based on Bayesian surprise was developed for calcification detection, while an Iris filter followed by a closed contour segmentation method made in the original coordinate system was used to detect and segment masses.

Calcifications and masses, when they exist, can be either benign or malign. BI-RADS describes important factors for malignancy determination including the distribution and morphology for calcifications and margins, shape and density for masses. A review of features used in the literature was performed and a lack of consensus on the adequate set of automatic features for the characterization of these findings was found. In this way, a large portion of the existing features were evaluated on the INbreast database by using the Pearson correlation, distance correlation and the Maximal Information Coefficient. These metrics were used to select appropriate subsets of features.

Based on the level of suspicion, lesions can be placed into one of seven BI-RADS categories: 0 when the exam is not conclusive, 1 for no findings, 2 for benign findings, 3 for probably benign findings, 4 for suspicious findings, 5 when there is a high probability of malignancy, and 6 for proven cancer. When more than one finding is present in the mammogram, the overall BI-RADS in the medical report corresponds to the finding with the highest BI-RADS. The typical learning settings described in the literature do not optimally represent this particular setting. In this way, a new learning paradigm is proposed, named max-ordinal learning (MOL), which sits in between supervised and semi-supervised classification. For every observation, some information about the label is available. However, in a subset of the examples, the knowledge is incomplete. This corresponds to the worst-case classification of the individual views of the example. A formalization of the max-ordinal learning paradigm led to two new learning schemes, MOL.LA and MOL.CD. MOL.CD uses coordinate descent in the space of the models, while in MOL.LA, the focus is on the partitioning of training instances into the two subsets. The experimental evaluation showed that the methodologies developed give better results than traditional approaches.

All the described techniques were thoroughly evaluated both independently (assuming all the previous step are correct) and in connection. It could be concluded that pectoral muscle detection, screening, mass segmentation, feature extraction and BI-RADS classification are ready to be used in practice. The calcification and mass detection algorithms, however, need to be improved in order to provide higher sensitivities with fewer false positives.

The impact of the conducted research will be reflected in its ability to improve the quality of breast cancer detection, speeding up the time to output a diagnosis with the correspondent beneficial implications in treatment possibilities and psychological patient well being. The radiologist will also benefit from the fact that he can better use his time concentrating on more difficult cases.

The author thanks Fundação para a Ciência e a Tecnologia (FCT) - Portugal for the financial support through the PhD grant with reference SFRH/BD/70713/2010.

Contents

List of Figures

List of Tables

Acronyms

ABUS	Automated whole Breast UltraSound
ACR	American College of Radiology
AD	Average Distance
ADI	Agência de Inovação
AMAE	Average Mean Absolute Error
AMED	Average Minimum Euclidean Distance
ANN	Artificial Neural Network
AOM	Area Overlap Measure
BB	Bounding Box
BI-RADS	Breast Imaging Reporting And Data System
BT	Breast Tomosynthesis
CAD	Computer-Aided Detection and Diagnosis
CaPTOR	Closed PaTh in Original cooRdinates
CC	CranioCaudal view
CD	Coordinate Descent
CEDM	Contrast-Enhanced Digital Mammography
CFS	Correlation-based Feature Selection
CM	Combined Measure of under-segmentation, over-segmentation and AOM
COCR	Cost-sensitive Ordinal Classification via Regression
corr	Pearson correlation
dcorr	distance correlation
DDSM	Digital Database for Screening Mammography
DICOM	Digital Imaging and COmmunications in Medicine
DM	Digital Mammography
EL	End-point on the Left column
ERR	Expected Reciprocal Rank
ET	End-point on the Top row
FDA	Food and Drug Administration
FFDM	Full Field Digital Mammography

FN	False Negative
FNr	False Negative rate
FP	False Positive
FPr	False Positive rate
GI	Gamma Imaging
GLCM	Grey-Level Co-Occurrence Matrix
GLRL	Grey-Level Run Length
GT	Ground Truth
HD	Hausdorff Distance
HGD	Histograms of Gradient Divergence
ICC	Intraclass Correlation Coefficient
KDLOR	Kernel Discriminant Learning Ordinal Regression
kNN	k-Nearest Neighbours
LA	Local Approximation
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
MAE	Mean Absolute Error
MIAS	Mammographic Image Analysis Society digital mammogram database
MIC	Maximal Information Coefficient
MIL	Multiple Instance Learning
MLO	MedioLateral Oblique view
MOL	Max-Ordinal Learning
MRI	Magnetic Resonance Imaging
PACS	Picture Archiving and Communication System
PSL	Partially Supervised Learning
RBF	Radial Basis Function
RG	Region Growing
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
RT	Reduction Techniques
SBC	Single Binary Classifier
SDF	Spatial Density Function
SFM	Screen Film Mammography
SMO	Sequential Minimal Optimization
SPLE	Shortest Path with Learnt Endpoints
SPPC	Shortest Path in Polar Coordinates
SSL	Semi-Supervised Learning
std	standard deviation
SVM	Support Vector Machine
SVR	Support Vector Regression
TN	True Negative
TNr	True Negative rate
TOR	Transductive Ordinal Regression
TP	True Positive
TPr	True Positive rate
US	UltraSound

Chapter 1

Introduction

Breast cancer is any form of malignant tumour which develops from breast cells. It is the most common cancer in women worldwide, and the leading cause of death from cancer in women, especially those between 40 and 55 years of age (Lan et al., 2012; Ganesan et al., 2013a; Hela et al., 2013).

In Portugal, 4 500 out of the 5 million female population are diagnosed with breast cancer every year, meaning that approximately 10% of Portuguese women will develop breast cancer at some stage of their lives. Each day 11 new cases are detected and 4 women will die (LPCC, 2009). Figure 1.1 depicts the incidence rates of breast cancer around the world.

Figure 1.1: Incidence rates of breast cancer worldwide (pink being the highest per capita rate) (Cancer, 2009).

Although some risk factors have already been identified such as age and family history of breast cancer (Nithya and Santhi, 2012), no study has yet been able to identify with certainty why every year one million breast cancer cases are discovered and 400 000 women worldwide die (Hela

et al., 2013). As a consequence, no primary prevention strategies have been identified. Early detection, however, can improve survival rates to a great extent (Tang et al., 2009; Ganesan et al., 2013a; Hela et al., 2013). Notwithstanding, no breast cancer symptoms appear at early stages (Nithya and Santhi, 2012). Several imaging techniques, including UltraSound (US) and Magnetic Resonance Imaging (MRI), have been proposed for the detection of breast cancer as early as possible. But mammography, a method that involves low-dose X-ray imaging of the breast, is currently the only widely accepted imaging method used for routine breast cancer screening (Tang et al., 2009; Ganesan et al., 2013a; Hela et al., 2013).

Screening mammography is performed in the asymptomatic population to detect early signs of breast cancer such as masses, calcifications, bilateral asymmetry and architectural distortion. Screening studies should be relatively low cost and present a high sensitivity: the screening test should pick up as much disease as possible, with the idea that subsequent studies will provide more specificity regarding the diagnosis. Diagnostic mammography is performed on patients who have already demonstrated abnormal clinical findings (Bozek et al., 2008; Tang et al., 2009; Ganesan et al., 2013a).

Both screening and diagnostic mammography are traditionally performed by radiologists who visually inspect mammograms (Ganesan et al., 2013a). This is not an easy task. One of the difficulties with mammography is that images generally have low contrast (Tang et al., 2009). Mammograms show normal structures such as fat, fibroglandular tissue, breast ducts and nipples, as well as possible abnormalities. Although fat appears as black regions on mammograms, everything else (glands, connective tissue and abnormalities) appear as levels of white, making it hard to distinguish between normal and abnormal tissue (Bozek et al., 2008; Hela et al., 2013). Furthermore, due to the advised two year interval screening routines for women between 45 and 69 years old (taking Portugal as an example), there is a large number of mammograms to be analysed every day. As a consequence, during manual screening, radiologists may get easily worn-out, missing vital cues while studying the scans. Manual screening is a tiring and tedious task prone to human error, for which the lack of a "gold standard" is an additional complicating factor.

Supporting these facts, studies have shown that mammography is susceptible to a high rate of False Positives (FPs) as well as False Negatives (FNs). Radiologists classify between 10% to 30% of malign cases as benign (Sampat et al., 2005b; Ganesan et al., 2013a). In these FN cases, the best time interval for the treatment of cancer can be missed, thus potentially endangering the patient. On the other hand, a high proportion of women without cancer (FP cases) undergo further unnecessary clinical evaluation or breast biopsy which can lead to needless anxiety. In either case, the costs associated with errors of misclassification are considerable, including emotional and economic burden, or even loss of life. To overcome these limitations, double reading of mammograms has been advocated. In fact, this is the procedure followed in every country that adopts a screening policy. The idea is to have two radiologists read the same mammographic images. This has been shown to increase sensitivity (equivalent to decreasing the number of FNs). However, the workload and cost associated with double reading remain high, and the outcomes are still susceptible to human error.

In the light of the difficulties involved in manual screening, the search for automated screening of mammograms or Computer-Aided Detection and Diagnosis (CAD) of breast cancer has been encouraged (Tang et al., 2009; Ganesan et al., 2013a). The present thesis describes an effort to develop automatic methods to help radiologists both in the screening and diagnosis phases.

Contributions and Document outline

After this introductory chapter, Chapter 2 presents the main background notions on breast cancer. Each of the following chapters of this document describe an incremental contribution in the field of computerized mammogram analyses. The pipeline of the adopted approach is shown in Figure 1.2 and a brief description is given as follows:

- Chapter 3: one of the main pre-processing steps when dealing with mammography is to detect the pectoral muscle region. Two new approaches, based on graph theory, are proposed and evaluated;

- Chapter 4: in this chapter, a fully automatic system to speed up the screening process is proposed which aims to detect normal breasts;

- Chapter 5: the two most common findings present in mammogram images are calcifications and masses. Due to their different characteristics, it is difficult to have a single detection method that works in both cases. The approaches used for the detection of each one of these findings are presented in this chapter;

- Chapter 6: in order to classify the images, it is important to summarize the characteristics of the findings into meaningful features. In this chapter, existing features are reviewed. Then, a large subset of features is tested and selected. The selected features will be used as input to the classification techniques proposed in the subsequent chapter;

- Chapter 7: this chapter introduces a new classification setting, motivated by the Breast Imaging Reporting And Data System (BI-RADS) standard. The problem is formalized and two different methodologies are proposed to solve it. Results on the INbreast database are also presented.

The results presented in each of the above mentioned chapters are independent of the previous task, typically assuming that all the preceding tasks were successful. For instance, when extracting features from masses, a perfect detection and segmentation steps are assumed by using the Ground Truth (GT) information. Chapter 8 presents an evaluation of the full integrated system in order to determine the block which affects the most the final overall performance, and thus in which steps of the pipeline future work should focus. The document ends in Chapter 9 with a summary of the contributions and directions for future work.

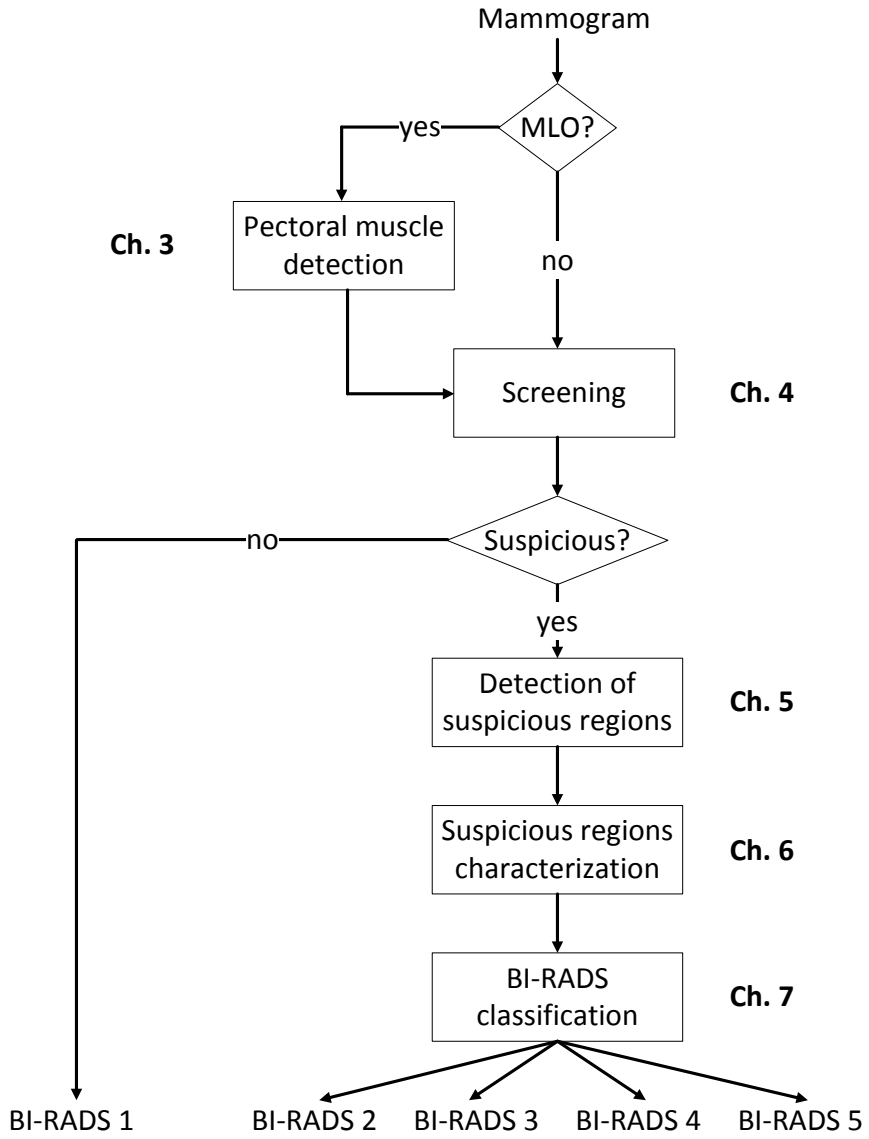


Figure 1.2: Pipeline of the adopted mammogram analysis approach. MLO stands for MedioLateral Oblique view.

Dissemination

My interest in the breast cancer field started with the work for the project:

- Advanced Objective Method for the Evaluation of the Aesthetic Result of Breast Interventions (FCT - PTDC/EIA/64914/2006) funded by the Portuguese Foundation for Science and Technology (FCT), with a budget of 95K EUR for 2007/2010 (`http://medicalresearch.inescporto.pt/breastresearch/index.php`).

I continued developing techniques for breast image processing and classification within the project:

- Semantic PACS (Picture Archiving and Communication System with Semantic Search Engine) funded by the Portuguese Agency for Innovation (ADI), with reference 3472, with a budget of 320K EUR for 2009/2011 (`http://www.inescporto.pt/~jsc/projects/SPACS/`).

I continued applying all the learned techniques in the context of the project:

- BCS project (Módulos Automáticos de Auxilio ao Rastreio e Diagnostico do Cancro da Mama Integrados em Sistemas PACS - project 33928), funded by the Portuguese Agency for Innovation (ADI), with a budget of 302K EUR for 2013/15.

The work developed in the context of my doctoral program has been published and presented as listed below.

- International Journals
  - J. S. Cardoso, I. Domingues, and H. P. Oliveira, "Closed Shortest Path in the Original Coordinates with an Application to Breast Cancer", International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), 2014.
  - E. Kozegar, M. Soryani, B. Minaei and I. Domingues, "Assessment of a novel mass detection algorithm in mammograms", Journal of cancer research and therapeutics, vol. 9, no. 4, 2013.
  - I. Domingues and J. S. Cardoso, "Max Ordinal Learning", IEEE Transactions on Neural Networks and Learning Systems, 2013.
  - I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Towards a Full Field Digital Mammographic Database", Academic Radiology, vol. 19, pp. 236-248, 2012.
- International Conferences
  - S. Bessa, I. Domingues, J. S. Cardoso P. Passarinho, P. Cardoso, V. Rodrigues and F. Lage "Normal breast identification in screening mammography: a study on 18 000 images", International Conference on BioInformatics and BioMedicine (BIBM), 2014, pp. 325-330.
  - I. Domingues and J. S. Cardoso "Using Bayesian surprise to detect calcifications in mammogram images", 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014.
  - J. S. Cardoso, R. Sousa, and I. Domingues, "Ordinal Data Classification Using Kernel Discriminant Analysis: A Comparison of Three Approaches", in International Conference on Machine Learning and Applications (ICMLA), 2012, pp. 473-477.
  - J. S. Cardoso and I. Domingues, "Max-Coupled Learning: Application To Breast Cancer", in International Conference on Machine Learning and Applications (ICMLA), 2011, pp. 13–18.

  - I. Domingues, J. S. Cardoso, I. Amaral, I. Moreira, P. Passarinho, J. S. Comba, R. Correia, and M. J. Cardoso, "Pectoral muscle detection in mammograms based on the shortest path with endpoints learnt by SVMs", in 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2010, pp. 3158-3161.
  - J. S. Cardoso, I. Domingues, I. Amaral, I. Moreira, P. Passarinho, J. S. Comba, R. Correia, and M. J. Cardoso, "Pectoral muscle detection in mammograms based on polar coordinates and the shortest path", in 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2010, pp. 4781-4784.

- National Conferences

  - I. Domingues and J. S. Cardoso, "Mass detection on mammogram images: A first assessment of deep learning techniques", in 19th Portuguese Conference on Pattern Recognition (RECPAD), 2013.
  - J. L. da Fonseca, J. S. Cardoso, and I. Domingues, "Pre-CADs in Breast Cancer", in 2nd PhD. Students Conference in Electrical and Computer Engineering (StudECE), 2013.
  - H. Zolfagharnasab, I. Domingues, and J. S. Cardoso, "Breast Density Classification: a comparison between Ordinal and Traditional Classification", in 2nd PhD. Students Conference in Electrical and Computer Engineering (StudECE), 2013.
  - I. Domingues, A. V. Alvarenga, and W. C. A. Pereira, "SVM classification of breast tumours on ultrasound images using morphological features", in XXIII Congresso Brasileiro de Engenharia Biomedica (CBEB), 2012.
  - I. Domingues, E. Sales, and W. C. A. Pereira, "INbreast-database masses characterization", in XXIII Congresso Brasileiro de Engenharia Biomedica (CBEB), 2012.
  - I. Domingues, J. S. Cardoso, and P. Cardoso, "Identification of benign breasts during mammogram screening", in 18th Portuguese Conference on Pattern Recognition (RECPAD), 2012.
  - I. Domingues and J. S. Cardoso, "Max Ordinal Learning Applied to the BI-RADS Classification of Mammograms", in 1st PhD. Students Conference in Electrical and Computer Engineering (StudECE), 2012.
  - I. Domingues and J. S. Cardoso, "Max-Coupled Ordinal Classification", in 17th Portuguese Conference on Pattern Recognition (RECPAD), 2011.
  - D. Gaspar, L. Bre, N. Bischoff, and I. Domingues, "Mammograms image processing", in Investigação Jovem na U.Porto (IJUP), Porto, 2011.
  - I. Domingues, J. S. Cardoso, and M. J. Cardoso, "Multi-Source Automatic Breast Cancer Diagnosis", in 16th Portuguese Conference on Pattern Recognition (RECPAD), 2010.

  - J. Tkaczuk, I. Domingues, and J. S. Cardoso, "Microcalcification Detection in Full Field Mammograms", in 16th Portuguese Conference on Pattern Recognition (REC-PAD), 2010.

- Invited talks

  - I. Domingues, "An automatic mammogram system: from screening to diagnosis", 2nd Workshop on computational methods for prevention and treatment of breast cancer, 2015.
  - I. Domingues, "Semantic PACS presentation", Turning Subjective Into Objective (TSIO), 2011.
  - I. Domingues, "Programa Informatico de leitura de mamografias Semantic PACS", Congresso Oncologia Integrativa, 2011.

- Reports

  - S. Bessa, I. Domingues, and J. S. Cardoso, "BCS report: State of the Art in Pre-CAD and CAD in Breast Cancer Screening and Diagnosis", FEUP, QREN: nr 33928, 2014.
  - I. Domingues, "Multi-Source Automatic Breast Cancer Diagnosis - intermediate report", INESC Porto / FEUP, 2011.
  - J. Tkaczuk, I. Domingues, J. S. Cardoso, I. Moreira, I. Amaral, M. T. Andrade, and E. Carrapatoso, "Semantic PACS Report Computer Aided Mammography - recent advances", INESC Porto / FEUP, 2011.
  - I. Domingues, L. Tesfaye, I. Amaral, J. S. Cardoso, and M. T. Andrade, "Semantic PACS Report Computer Aided Mammography - a follow up", INESC Porto / FEUP, 2010.

I have also co-supervised the MSc thesis:

- J. C. L. da Fonseca, "Pre-CADs in Breast Cancer", MSc, Faculdade de Engenharia da Universidade do Porto, 2013.

and informally collaborated in two other MSc thesis, namely:

- C. M. C. C. Castro, "Estudo do impacto da densidade mamaria no cancro da mama", MSc, Faculdade de Medicina da Universidade do Porto, 2013.

- J. P. da S. F. Monteiro, "Computer aided detection in mammography", MSc, Faculdade de Engenharia da Universidade do Porto, 2011.

Conclusion

The impact of the conducted research aims to improve the quality of breast cancer detection, speeding up the time to output a diagnosis with the correspondent beneficial implications in treatment possibilities and psychological patient well being. The radiologist will also benefit from the fact that he can spend more time analysing the more challenging cases.

Among the contributions some are worth underlying. Two new pectoral muscle segmentation techniques, both based on the shortest path idea, have been proposed and validated. A new learning problem was motivated, formalized and solutions proposed. This framework was inspired by a specific application in the breast cancer field, but its use in numerous other applications is foreseen. Finally, an evaluation of the complete integrated system (from screening to diagnosis) is presented, where both the strengths and vulnerabilities are identified, leading to future work directions.

Chapter 2

Background and related work

The focus of this chapter is on the basic notions needed to fully understand the context of the thesis[1]. It starts with Section 2.1 where the basics of breast anatomy are illustrated. Breast cancer is presented in Section 2.2. The main existing breast image techniques are compared in Section 2.3 and in Section 2.4 some detail is given on mammography, the technique that provides the images studied in this dissertation. Section 2.5 describes the existing mammogram databases and justifies the creation of a new database, presented in Section 2.6. Some of the most widespread commercial Computer-Aided Detection and Diagnosis (CAD) systems are mentioned in Section 2.7. The adopted evaluation methodologies are clarified in Section 2.8. Finally, Section 2.9 draws conclusions from this chapter.

2.1 Normal breast anatomy

The breasts are a secondary sexual characteristic of the female reproductive system whose primary roles are related to sexual attraction and to provide milk for the nourishment of the infant, Figure 2.1. They are located on the anterior and lateral parts of the chest, overlying the chest (pectoral) muscles, Figure 2.2.

The two parts of the breast are the interior or glandular part and the exterior. The exterior part includes the nipple, areola and tubercles (small elevations in the areola). The interior part, which is also the principal secretory organ, is assembled in 15 to 25 lobes of compound milk-producing glands embedded in fibrous and adipose tissue. Each of these lobes contains an excretory duct that drains the lactiferous sinus (Chhaya et al., 2013).

Fat and other tissue fills the spaces between the lobes and ducts. The breasts also contain lymph vessels, which are connected to small, round masses of tissue called lymph nodes. Lymph nodes produce cells that help the body fight infections. Groups of lymph nodes are near the breast in the underarm, above the collarbone, and the chest behind the breastbone (NIH, 2014).

[1] Publications related with this chapter include (Moreira et al., 2012; Domingues et al., 2012c).

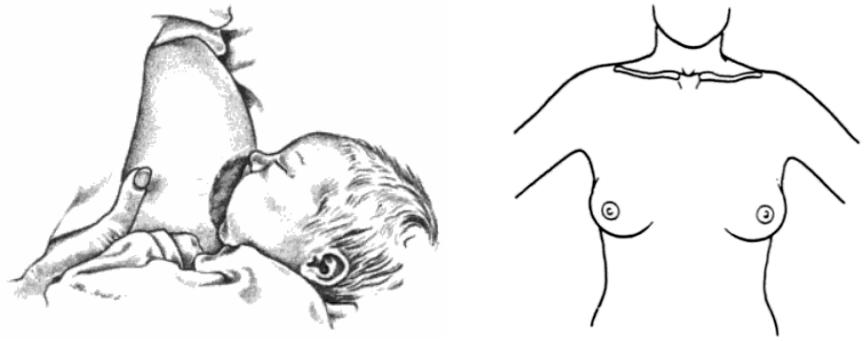


Figure 2.1: Normal breasts function and anatomy. Left: Breastfeeding (Cadwell et al., 2006); Right: Breasts location (Hindle, 1999).

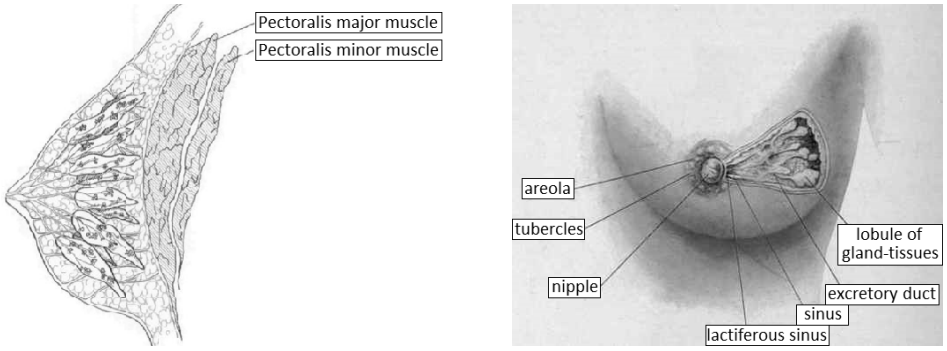


Figure 2.2: Normal breast structure. Left: pectoral muscle (Morris and Liberman, 2005); Right: secreting structure of the breast (Davis, 1913).

2.2 Breast cancer

Breast cancer is not a single disease but is highly heterogeneous at both the molecular and clinical level. Tumours originate from a single cell that acquired multiple mutations and unlimited proliferative potential. In this way, the natural history of breast cancer involves progression through defined pathological and clinical stages, starting with ductal hyper-proliferation, subsequent evolution into *in situ* and invasive carcinomas, and finally into metastatic disease (Figure 2.3).

Breast tumours are classified histologically according to the location of the origin, Figure 2.4. The ductal tumours develop in breast ducts and represent 80% of the tumours. The lobular tumours develop inside the lobes and account for 10 to 15% of cases. Other subtypes represent fewer than 10% of cases diagnosed per year (Hirata et al., 2014).

Breast tumours can be benign (not cancer) or malignant (cancer). Benign tumours are rarely a threat to life. They can be removed and usually do not grow back. They also do not invade the tissues around them and do not spread to other parts of the body. Malignant tumours, on the other side, can be a threat to life. They can often be removed but sometimes grow back. They can also invade and damage nearby organs and tissues (such as the chest wall) and can spread to other parts of the body (NIH, 2014).

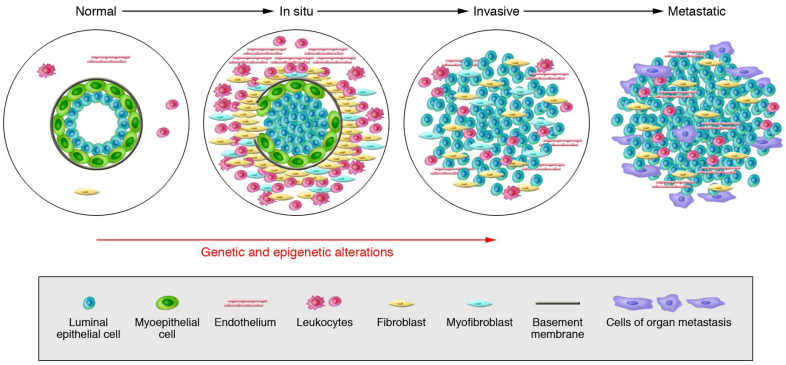


Figure 2.3: Hypothetical model of breast tumour progression (Polyak, 2007).

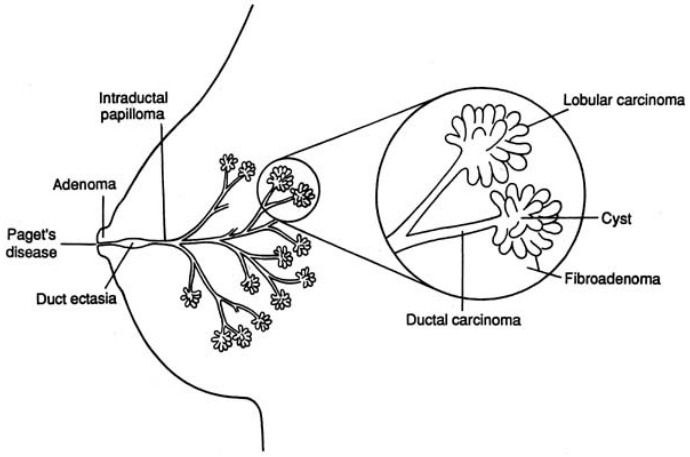


Figure 2.4: Diagram of the common types and sites of breast pathology (Hindle, 1999).

Lumps that are soft, smooth, round and movable are likely to be benign. Hard, oddly shaped ones that feel firmly attached within the breasts are more likely to be cancerous (NIH, 2014). To make a diagnosis, imaging techniques are used.

2.3 Breast imaging

Breast imaging is continuously progressing as new modalities, and advances in applications of established modalities, are developed. A summary of the most commonly used techniques is presented in Table 2.1, along with their most significant strengths and weaknesses. Some comparative images are also shown in Figures 2.5, 2.6, 2.7, 2.8, 2.9, 2.10, and 2.11.

Although each technique has its advantages and pitfalls, mammography is still the only widely adopted screening technique. While traditionally Screen Film Mammography (SFM) was used, nowadays most medical institutions have access to Digital Mammography (DM). In the next section, a brief overview of the image acquisition method and typical findings in mammography is given.

Table 2.1: Breast imaging techniques: description and potential strengths and limitations (Vainio and Bianchini, 2002; Scaperrotta et al., 2008; Dromain et al., 2009; Kim et al., 2014a; Heywang-Koebrunner and Schreer, 2014).

Type	Technique	Potential strengths	Current limitations
Mammography	Screen Film Mammography (SFM), Figure 2.8	still widely used for screening; public databases	dense breasts
	Digital Mammography (DM) Figure 2.8	easy display, transmission and storage	higher cost than SFM
	Contrast-Enhanced Digital Mammography (CEDM), Figure 2.7	extent of the disease; response to therapy	intra-venous injection of an iodinated contrast agent
UltraSound	UltraSound (US), Figure 2.11	increased sensitivity for dense breasts; no radiation	operator-dependent; more expensive and less specific than mammography
	Sonoelastography	mechanical properties of tissues	high FNr
	Automated whole Breast UltraSound (ABUS), Figure 2.5	low interobserver variability	low sensitivity; high FPr
Tomography	Breast Tomosynthesis (BT), Figure 2.6	dense breasts	increased radiologist reading time
	Positron Emission Tomography	staging of breast cancer	expensive; limited access; low sensitivity
Magnetic Resonance Imaging (MRI)	MRI, Figure 2.10	more sensitivity than mammography; no radiation	more expensive and less specific than mammography; patient claustrophobia
Gamma Imaging (GI)	Scintimammography, Figure 2.9	dense breasts	radiation; poor spatial resolution

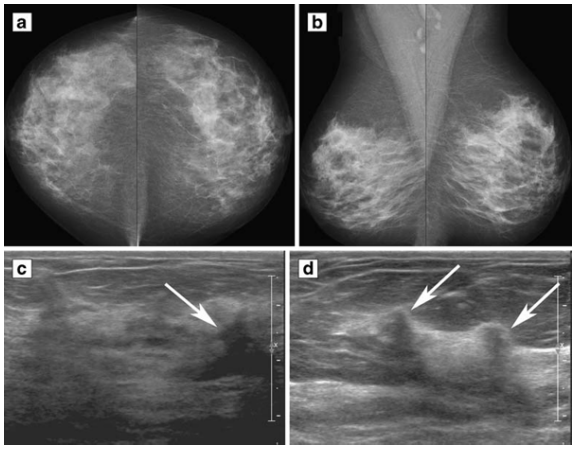


Figure 2.5: 54-year-old asymptomatic woman with dense breasts and no previous history of breast cancer. a) CranioCaudal (CC) DM; b) MedioLateral Oblique (MLO) DM; c) Transverse ABUS image of the right breast (white arrow shows an invasive ductal carcinoma); d) ABUS image of the left breast (two white arrows show an invasive carcinoma with lobular carcinoma in situ) (Kelly et al., 2010).

2.4 Mammography

Mammography is a medical imaging technique that is based on x-ray examinations of the breast. The resulting x-ray images are called mammograms. Typically, the breast is imaged from two views: MLO, a side/angled view, and CC, a view from above (Figure 2.12). This procedure has proved to be an effective tool in reducing the mortality due to breast cancer (Naqa and Yang,

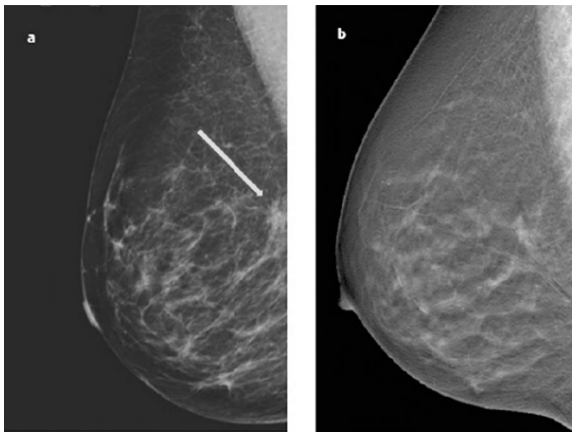


Figure 2.6: Mammography versus BT. Left: Mammogram; Right: BT. While the mammogram image shows a possible spiculate mass (arrow), the BT view demonstrates this was due to superimposition of glandular tissue (Heywang-Koebrunner and Schreer, 2014).

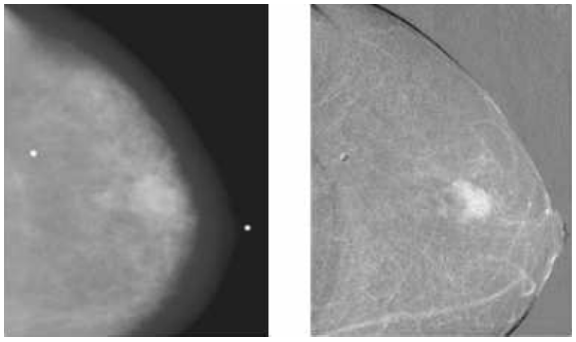


Figure 2.7: Example of (a) pre-contrast image and (b) post-contrast image for a simple model of digital subtraction with dual energy mammography (Sanchez, 2005).

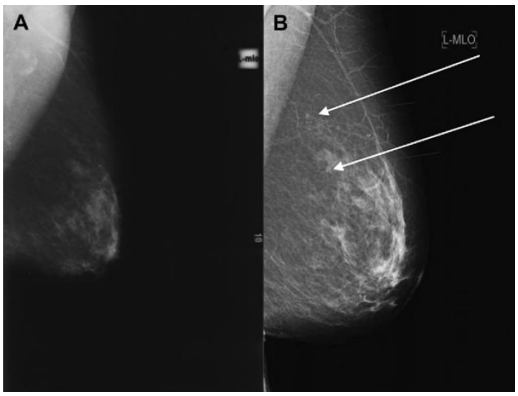


Figure 2.8: SFM versus DM. Patient with SFM called back for asymmetric density seen on CC view only, which proved to be normal tissue. A significant finding on DM (B) of two 5-mm invasive ductal carcinomas (arrows) not seen on SFM (A) imaging (Green and Weiss, 2013).

2005).

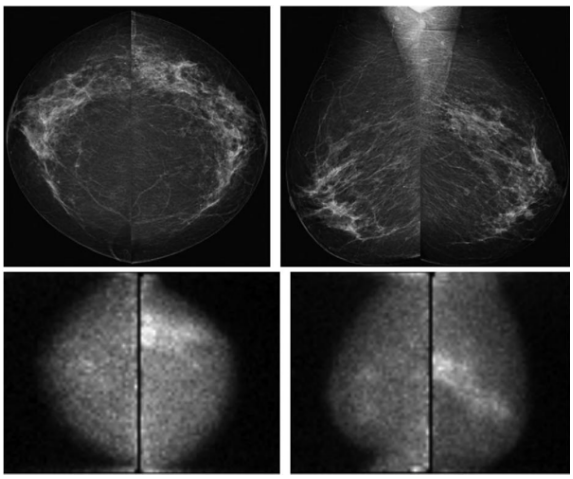


Figure 2.9: Dense breast: Top: DM; Bottom: GI image of the same patient showing ductal carcinoma *in situ* (Green and Weiss, 2013).

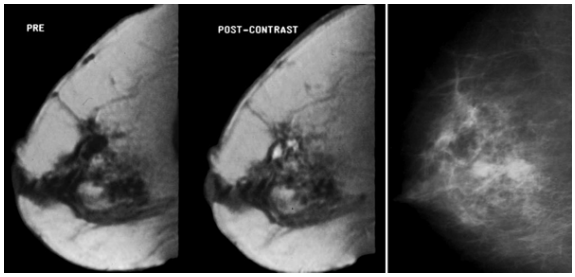


Figure 2.10: Mammographically occult carcinoma in a dense breast. MRI (left) are obtained before (PRE) and after (POST) gadolinium administration. The white area in the post-contrast image denotes marked enhancement of a centrally located breast carcinoma. A mammogram of the same breast (right) shows dense tissue, but no lesion at the location where the cancer was detected by MRI (Conant and Maidment, 1996).

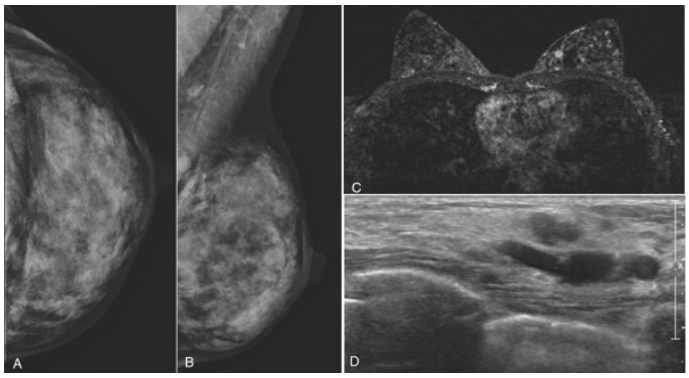


Figure 2.11: High risk 53-year-old patient breast images. A and B: normal left mammogram; C: MRI at 6-month interval revealed a 7-mm mass; D: Targeted USG finds a left 9 o'clock abnormality (Kuhl and Mahoney, 2013).

Figure 2.12: Common views in mammography: CC and MLO views from the same patient in both breasts (Sanchez, 2005).

Mammography shows the morphological aspects of the breast, such as the anatomical structures and all the breast tissues, namely glandular, fibrous and adipose tissues. There is, however, one major problem concerning breast radiography: these three types of tissue have similar density among them (there is no bone tissue or air to provide high contrast). The relative density of the breast is primarily affected by the inherent characteristics of the patient, e.g. their hormonal status, age and pregnancy (Moreira, 2012). The American College of Radiology (ACR) classifies breast density according to the classes in Table 2.2 (Garcia-Manso et al., 2013). These differences are visible in the examples of Figure 2.13 where four sketches with different ACR classes are shown.

Table 2.2: ACR categories.

Category	Description	
I	almost no fibroglandular tissue	$< 25\%$ glandular
II	scattered fibroglandular densities	$\approx 25-50\%$ glandular
III	heterogeneously dense	$\approx 51-75\%$ glandular
IV	extremely dense	$> 75\%$ glandular

(a) ACR I **(b)** ACR II **(c)** ACR III **(d)** ACR IV

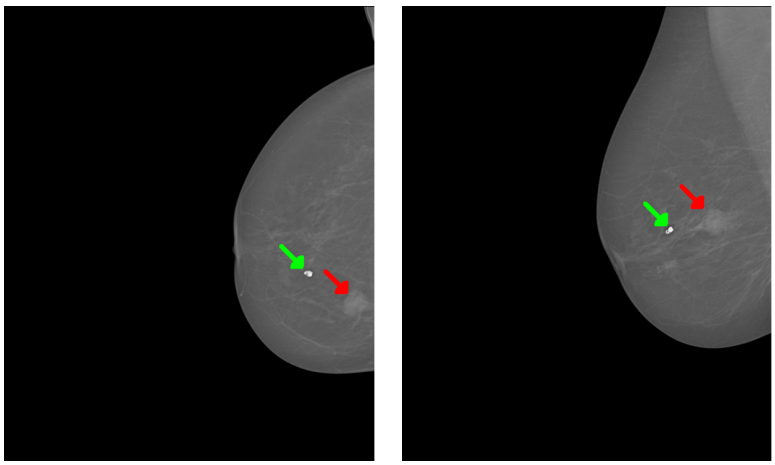
Figure 2.13: ACR categories exemplification (ACR, 2014).

There are mainly two types of examinations performed using mammography: screening mammography and diagnostic mammography. Screening mammography is performed to detect breast

cancer in an asymptomatic population. The aim of diagnostic mammography is to examine a patient who has already demonstrated abnormal clinical findings. Diagnostic mammography is often performed as a follow up examination of an abnormal screening mammography in order to determine whether the area of concern on the screening examination needs additional breast imaging or a biopsy to determine whether the woman has breast cancer. The adoption of mammographic examinations, especially screening mammography, has been proven to increase the rate of detection of cancer and to reduce the rates of morbidity and mortality due to earlier phases of diagnosis (Tang et al., 2009; Helvie et al., 2014).

The major drawback of projection radiography is that X-ray beams project the original anatomical three-dimensional objects onto a two-dimensional image (Lo et al., 1998). In other words, each pixel intensity on the image represents a total X-ray attenuation integrated from a line passing through the patient. Soft tissue and abnormal changes of tissue can be distinguished in X-ray because they attenuate X-rays differently. However, subtle anomalies superimposed on various normal tissues are difficult to discern.

The most common radiologic feature of ductal cancer *in situ* is calcifications, while the most common radiologic feature of lobular cancer is named mass. An example of a breast presenting both findings is shown in Figure 2.14.



(a) CC view **(b)** MLO view

Figure 2.14: Breast exam presenting multiple findings. Green arrows point to a calcification, red arrows point to a mass.

According to the National Cancer Institute, calcifications are deposits of calcium in the tissues. Benign calcifications are usually larger, coarser, round, with smooth margins, and are much more easily seen than calcifications associated with malignancy. Calcifications associated with malignancy are usually very small. Many calcifications clustered together may also be a sign of cancer.

According to ACR, a mass is a three-dimensional structure demonstrating convex outward borders, usually evident on two orthogonal views. It may be caused by the abnormal growth of

cells, a cyst, hormonal changes, or an immune reaction.

Both masses and calcifications may be benign (not cancer) or malignant (cancer). An exam can have only calcifications, only masses, or both. When reporting an exam, the specialist is advised by ACR to follow the Breast Imaging Reporting And Data System (BI-RADS) standard. The standard is composed of 7 classes, summarized in Table 2.3.

Table 2.3: BI-RADS categories, interpretation and recommended actions.

Category	Interpretation	Recommended action
0	Insufficient study	Obtain additional imaging
1	Negative	Routine follow-up
2	Benign finding(s)	Routine follow-up
3	Probably benign finding(s)	Short-interval follow-up
4	Suspicious finding	Biopsy should be considered
5	Highly suggestive of malignancy	Biopsy necessary
6	Biopsy proven malignancy	

Since BI-RADS classification summarizes the full exam in a standardized way, one of the main aims of this thesis is to develop techniques which can make a reliable automatic classification of the BI-RADS category given the mammograms.

2.5 Available databases

There are several mammogram databases, some public and some restricted to individual groups, which are used by researchers in the breast cancer area. However, these often do not meet all the requirements needed for a study. The following paragraphs describe existing databases and summarize their main limitations.

The Mammographic Image Analysis Society (MIAS) digital mammogram database (Suckling et al., 1994), despite being the oldest available database, is still in use. It consists of 161 cases, 322 digitized MLO images, with all types of findings, including benign, malign and normal images. It has a high percentage of spiculated masses but Rangayyan et al. (2000) noticed that there was a disproportionate number of benign findings in relation to the malign cases. The dataset contains breast density information though it is not classified according to ACR standards. MIAS annotations consist of the centre and radius of a circle around the area of interest. These types of annotations are not considered sufficient for some studies, such as the one by Oliver et al. (2010), where all circumscribed and spiculated lesions had to be manually segmented. Another drawback is the resolution to which the images have been digitized, which makes MIAS unsuitable for experiments on the detection of calcifications (Dominguez and Nandi, 2007).

The most used database is the Digital Database for Screening Mammography (DDSM) (Heath et al., 1998). It is the largest public database, with 2 620 cases including 2 images from each breast (MLO and CC), making a total of 10 480 images, with all types of findings from normal images to images with benign and malign lesions. Some of the cases in this database were collected from the Nijmegen Database (Netsch and Peitgen, 1999). The only patient information included is the age, but it has breast density annotations (ACR) and BI-RADS annotations. Image annotations include pixel level boundary of the findings. However, as noted in (Singh and Bovis, 2005; Sampat et al.,

2008; Song et al., 2010), DDSM annotations are not adequate for the validation of segmentation algorithms because the precision of the annotations is not high enough.

A more recent database is the "BCDR: a breast cancer digital repository" (Lopez et al., 2012). The BCDR is subdivided in two different repositories: (1) a Film Mammography-based Repository (BCDR-FM) and (2) a Full Field Digital Mammography (FFDM) Repository (BCDR-DM). The BCDR-FM is composed of 1 010 (998 female and 12 male) patients cases, including 1 125 studies, 3 703 MLO and CC mammography incidences and 1 044 identified lesions clinically described. The BCDR-DM, still in construction, is composed of 79 biopsy-proven lesions of 64 women, rendering 143 segmentations an average of 1.81 images per lesion, including clinical data and image-based descriptors (Moura and Lopez, 2013). The raw images (CC and MLO) are also available together with the coordinates of the lesion's contours and numerical anonymous identifiers for linking instances and lesions. It is a binary class dataset due to the initial BI-RADS classification of the radiologist being replaced by the result of the biopsy (Benign vs. Malign finding).

As can be seen, both MIAS and DDSM have several limitations, the most important fact being that they are composed by digitized (i.e. scanned) images and not digital ones. The digitization process can introduce several artefacts that are no longer problematic with the event of fully digital mammogram images. At the time of writing this thesis, the FFDM part of BCDR contains a limited number of cases. An FFDM database was thus built and its description is given in Section 2.6.

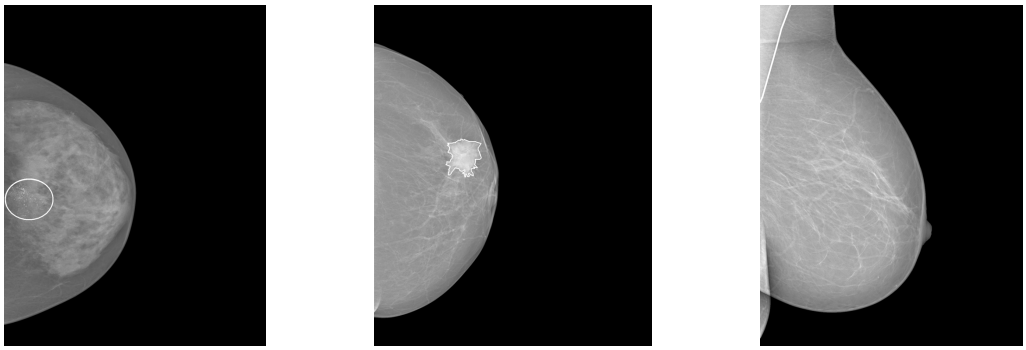
2.6 INbreast database

The database was collected at the Breast Centre of Hospital São João, Porto, under permission of the Hospital's ethics committee. Images were acquired between April 2008 and July 2010. The acquisition equipment was the MammoNovation Siemens, with a Solid-state detector of amorphous selenium, with a pixel size of $70\mu m$ (microns), and 14bit contrast resolution. The image matrix was 3328×4084 or 2560×3328 pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). Images were saved in the Digital Imaging and COmmunications in Medicine (DICOM) format. All confidential information was removed from the DICOM file, according to Supplement 55 of the DICOM standard. The correspondence between images of the same patient is kept with a randomly generated Patient ID.

INbreast has FFDM images from screening, diagnostic and follow-up cases. Screening is made according to Portuguese national and regional standards (Lee et al., 2010). Diagnosis is made when the screening shows signs of anomaly. In follow-up images, cancer was previously detected and treated. A total of 115 cases were collected, of which 90 have 2 images (MLO and CC) of each breast and the remaining 25 cases are from women who had a mastectomy, and thus only 2 views of one breast were included. This sums to a total of 410 images. 8 of the 90 cases with 2 images per breast also have images acquired in different timings (follow-up).

The database includes examples of normal mammograms, mammograms with masses, mammograms with calcifications, architectural distortions, asymmetries, and images with multiple findings (Figure 2.14).

The main contribution of this database is the carefully associated Ground Truth (GT) annotation. The annotations were made by a specialist radiologist in the field, and validated by a second specialist. When there was a disagreement between the experts, the case was discussed until a consensus was obtained. Annotations were made on OsiriX, an open source Picture Archiving and Communication System (PACS) workstation, running on a Macintosh platform. Each finding has a label that identifies the type of lesion. There are 7 types of annotations: "Asymmetry", "Calcification", "Cluster" (of calcifications), "Distortion", "Mass", "Pectoral Muscle" (only in the MLO view), and "Spiculated Region". For the types, "Asymmetry", "Calcification", "Distortion", "Mass" (Figure 2.15), and "Pectoral Muscle" (Figure 2.15), a detailed contour of the finding was made. An ellipse enclosing the entire cluster was adopted to annotate the clusters of calcifications (Figure 2.15). When the mass is spiculated, in addition to the contour of the denser region, an ellipse enclosing all the spicules was added. To the best of our knowledge, INbreast is the only available database that contains information on the pectoral muscle contour. This information is important in several applications including the localization of the findings and registration of breast images.



(a) cluster of calcifications **(b)** masses **(c)** pectoral muscle

Figure 2.15: Annotation examples.

There are a total of 116 masses from 107 images (≈ 1.1 masses per image). The mean mass size is 479 mm^2 (with a *std* of 619 mm^2). The smallest mass has 15 mm^2 and the biggest has an area of 3689 mm^2.

Concerning calcifications, they are present in 301 of the 410 images. The tag "Cluster" was only used in 27 sets of calcifications, in 21 images (≈ 1.3 clusters per image). Of these 21 images, only 2 had no individual calcification annotation. A total of 6 880 calcifications were individually identified in 299 images (≈ 23.0 calcifications per image).

The annotations were saved in XML format with the following structure:

- A standard header with the XML version and type of encoding information;

- The tag <key>NumberOfROIs< /key> followed by an integer that indicates the number of annotations present in the image;
- For each Region Of Interest (ROI), there is a tag <key>Area< /key> followed by the value of the area of the current ROI, the tag <key>Centre< /key> followed by the coordinates of the point in the centre of the ROI, the tag <key>Name< /key> followed by the type of finding (Mass, Calcification, Distortion, Spiculated Region) and some other general information about the ROI;
- After the general information, for each ROI, a list of contour points is presented between the tags <array> and < /array>.

Information regarding the patient's age at the time of image acquisition, family history, ACR breast density annotation and BI-RADS classification is also provided. A biopsy result is also displayed whenever performed. A biopsy was performed on 56 cases of which 11 were found to be benign and the remaining 45 were malignant.

One of the most important breast characteristics is density. Dense breasts are harder to analyse through mammography than non-dense breasts (Elshinawy, 2010). For each image in INbreast, its density according to the ACR standard scale is available. A distribution of density for each BI-RADS class is presented in Figure 2.16.

With the precise annotations in INbreast, studies can be developed that cannot be performed with the other currently available databases. Shape information is highly indicative of the malignancy of the mass (Sun et al., 2010b) and therefore automatic shape assessment in the mammogram is often pursued. However, the coarse-grained annotation of current databases does not allow a proper validation of the discoveries. Also, the calcifications grouping and distribution in the mammogram is important for the correct diagnosis. Again, the usual annotation of the calcifications with a single region enclosing all calcifications is insufficient for the development of automatic methods. The database is available at `http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database` and has already been cited in studies including (Borges et al., 2014; Cardoso et al., 2014; Matheus et al., 2014; Wiemker et al., 2014; Barde and Sawarkar, 2013; Canul-Reich and Mendez, 2013; Domingues and Cardoso, 2013b; Fonseca, 2013; Fonseca et al., 2013; Kozegar et al., 2013; Matheus et al., 2013b,a; Moura et al., 2013; Moura and Lopez, 2013; Oliveira, 2013; Oliveira et al., 2013; Ortega et al., 2013; Suarez-Ortega and Franco-Valiente, 2013; Teixeira, 2013; Domingues et al., 2012b,c; Domingues and Cardoso, 2012; Graven and Samuelsen, 2012; Lopez et al., 2012; Moreira, 2012; Cardoso and Domingues, 2011; Domingues, 2011).

2.7 Commercial CAD systems

To date, the Food and Drug Administration (FDA) in the United States has approved at least three of the commercially available CAD systems developed to aid radiologists in detecting mammographic abnormalities (Sampat et al., 2005b). The first CAD approved by FDA was ImageChecker

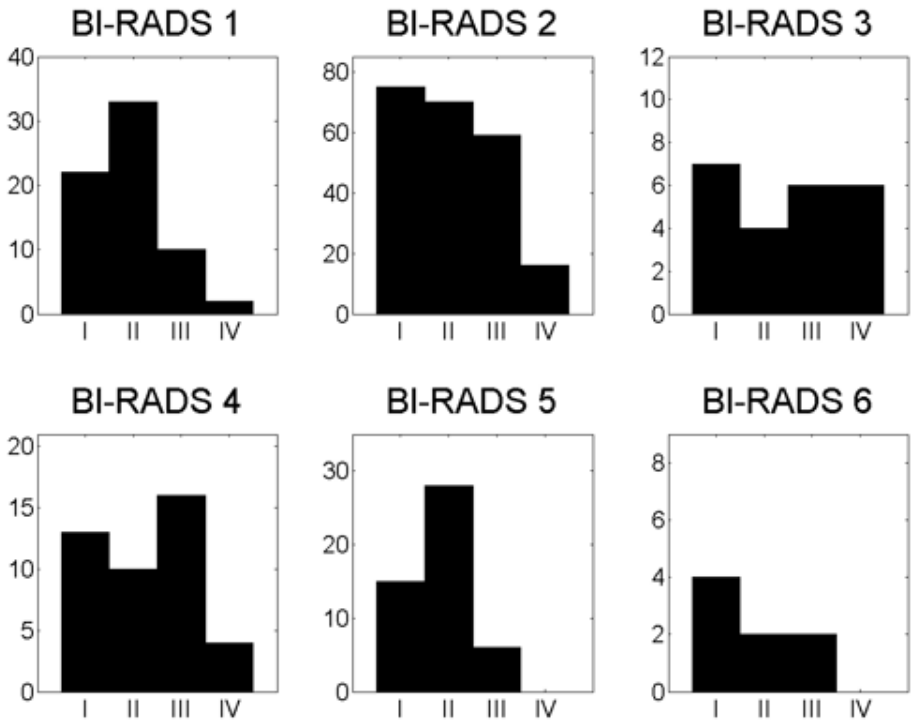


Figure 2.16: Distribution of density across BI-RADS scale. Horizontal axis corresponds to ACR density and vertical axis refers to the number of images.

of Hologic R2. Later, two new mammographic CAD systems were approved: MammoReader from iCad and SecondLook from CADx. Due to legal disputes, MammoReader was discontinued, and the product is now named SecondLook and is commercialized by iCad.

Since the techniques used by commercial vendors are proprietary, it is not possible to determine which algorithms each system uses. Some insight (that may be outdated) is however given in the NHSBSP report (NHSBSP, 2001). The ImageChecker takes advantage of algorithms originally developed by Karssemeijer and Brake (1998). It uses an "annulus", consisting of inner and outer concentric circles, which is moved over the image. When the software encounters line segments radiating from a common origin (the inner annulus) it marks the region of maximum convergence with an asterisk on the low resolution image. It is also sensitive to the presence or absence of a central density.

Normal breast structures such as benign calcified tissues or crossing linear tissues may be shown as false calcification prompts. False mass prompts include ducts and tissue radiating from the nipple and inadvertent scanning of parenchymal tissue. Manufacturers state that there is no systematic pattern of features likely to result in false negatives other than positioning, poor contrast or motion. Specificity is claimed to be 60% and the average number of false prompts per film is 0.6. The ImageChecker makes no diagnostic interpretation of the ROI.

For the SecondLook system, the NHSBSP report (NHSBSP, 2001) states that it uses similar

artificial intelligence and Artificial Neural Network (ANN) algorithms to those employed by US Air Force fighter jets. The product documentation claims to "detect up to 72% of actionable missed cancers an average of 15 months earlier than screening mammography alone" and to have "90 – 96% sensitivity with 2.0 or 2.9 false positives per 4-view study".

A collection of selected published assessments of these CAD systems is presented in Table 2.5 for calcifications, Table 2.4 for masses and Table 2.6 for an overall analysis. The same information is summarized in Figure 2.17.

Table 2.4: Independent evaluation of CAD systems: calcifications

System	Reference	Year	Sensitivity (%)	FPr
ImageChecker	(Brem and Schoonjans, 2001)	2001	98	1
	(Baum et al., 2002)	2002	89	0.35
	(Soo et al., 2005)	2005	51	2.0
	(Sampat et al., 2005b)	2005	98.5	0.75
	(Masala, 2006)	2006	98	2.22
	(Cruz, 2011)	2011	98.5	0.74
	(Raman et al., 2011)	2011	98.5	0.74
SecondLook	(Malich et al., 2001)	2001	98.2	0.28
	(Malich et al., 2003)	2003	93.0	0.20
	(Brem et al., 2005a)	2005	95	0.65
	(Lobbes et al., 2013)	2013	58.1	0.88

Table 2.5: Independent evaluation of CAD systems: masses

System	Reference	Year	Sensitivity (%)	FPr
ImageChecker	(Vyborny et al., 2000)	2000	86	0.24
	(Baum et al., 2002)	2002	81	0.26
	(Sampat et al., 2005b)	2005	85.7	1.32
	(Masala, 2006)	2006	88	2.22
	(Ellis et al., 2007)	2007	81.8	1.08
	(Raman et al., 2011)	2011	85.7	1.32
SecondLook	(Malich et al., 2001)	2001	88.7	0.97
	(Malich et al., 2003)	2003	90.0	0.81
	(Brem et al., 2005a)	2005	87	2.30
	(Ellis et al., 2007)	2007	60.9	1.41
	(Lobbes et al., 2013)	2013	81.1	1.68

Table 2.6: Independent evaluation of CAD systems: all types of findings

System	Reference	Year	Sensitivity (%)	FPr
ImageChecker	(Moberg et al., 2001)	2001	22	0.482
	(Baum et al., 2002)	2002	87.3	0.61
	(Ikeda et al., 2004)	2004	42	2
	(Samulski et al., 2010)	2010	80.4	2.0
	(Hupse et al., 2013)	2013	75	0.21
MammoReader	(Sampat et al., 2005b)	2005	89.3	2.32
SecondLook	(Malich et al., 2001)	2001	90.0	1.3
	(Brem et al., 2005b)	2005	89	1.3
	(Brem et al., 2005c)	2005	89	1.3
	(Brem et al., 2005a)	2005	89	2.95
	(The et al., 2009)	2009	94	2.3
	(Lobbes et al., 2013)	2013	80.3	2.56

As usual, care must be taken when interpreting these results. Besides differences in the databases and in software versions, some results refer to the use of the CAD system alone, while others evaluate the performance of the specialist aided by the CAD system. Moreover, some

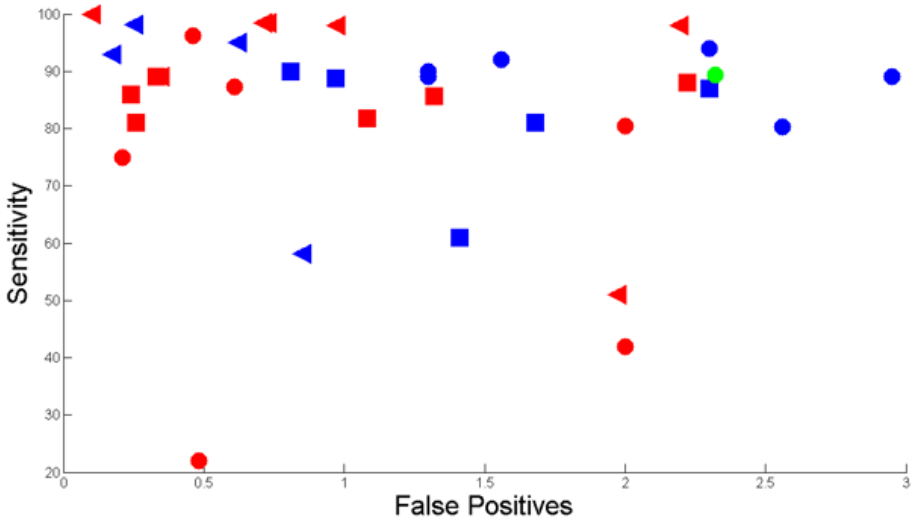


Figure 2.17: Independent evaluations of CAD systems. The triangles indicate calcifications, squares masses and circles all types of findings. Results using ImageChecker are plotted in red, MammoReader in green and SecondLook in blue. The aim of each system is to maximize sensitivity while minimizing the number of FPs (i.e. the region in the top left of the plot).

systems evaluate results per image, others per set of 2 views and yet more per case of 4 views. Nevertheless it can be concluded that calcification detection has better results than mass detection. This is an expected result since, as it has been noted by several authors, e.g. (Rangayyan et al., 2007), calcifications are easier to detect than masses. When comparing the existing systems, ImageChecker seems to behave better than SecondLook or MammoReader. The fact that there are more published results using ImageChecker than SecondLook or MammoReader reflects the probable higher usage of ImageChecker in real clinical practice.

Rangayyan et al. (2007) compiled some strengths and weaknesses of breast CAD. Although it is a work from 2007, some interesting points are:

- With the introduction of direct digital imaging systems to mammography (with increased contrast, dynamic range, and signal-to-noise-ration) there may no longer remain the need to image enhancement.
- Selective analysis of breast density can lead to improvements in the prediction of the risk of development of breast cancer based upon screening mammograms.
- The problem of calcification detection can be considered satisfactorily solved.
- There is a need to increase the sensitivity of detection of masses to higher values around 95% at low false positive rates of less than one per image. It is also desirable to indicate the degree of suspicion for each region identified.
- There exists the need for the development of improved detection and analysis techniques for accurate discrimination of spicules against blood vessels and ducts.

- More methods are desirable to analyse asymmetry from multiple perspectives, including pattern asymmetry in the fibroglandular tissue as well as morphological and density measures related to the breast and the fibroglandular disk.

- Accurate detection of architectural distortion can be the key to efficient detection of early breast cancer, at pre-mass-formation stages.

- Simultaneous analysis of current and prior mammograms can enhance the performance of CAD systems in the detection of breast cancer.

- Indexed atlases can be developed to help in the teaching and training of radiologists, and combined with content-based retrieval tools to help radiologists in the decision-making process for difficult-to-diagnose cases.

It is too large a problem to address all the open issues in the field in a single thesis. Therefore, only some of the above points are studied, namely to use digital (and not digitized) images which eliminates the need for image enhancement (Chapter 3), to develop screening methods that take breast density into consideration (Chapter 4), to verify if calcification detection is a solved problem and improve mass detection (Chapter 5) and to develop techniques to help the specialist make a diagnostic decision in the difficult-to-diagnose cases (Chapters 6 and 7). The study of spiculation, asymmetry, architectural distortion and temporal changes (among others) are left as future work, since it is too large of a problem to address all the open issues.

2.8 Evaluation Methodology

For a better understanding of the present document, each chapter includes, not only the methodology, but also a state of the art description and the evaluation of the proposed methodology. As the evaluation techniques apply to more than one chapter, they are summarized in the following paragraphs.

Segmentation results are evaluated using both Region and Contour-based metrics. The first includes the Area Overlap Measure (AOM), and a combined measure (CM) of under-segmentation (U), over-segmentation (O) and AOM (Elter et al., 2010) defined as:

$$AOM = \frac{|S \bigcap T|}{|S \bigcup T|} \qquad CM = \frac{AOM + (1-U) + (1-O)}{3}$$

where

$$U = \frac{|T \backslash (S \bigcap T)|}{|T|} \qquad O = \frac{|S \backslash (S \bigcap T)|}{|S|}$$

S represents the automatically segmented region, and T represents the GT region of the same lesion.

Contour-based metrics include Average Distance (AD), Average Minimum Euclidean Distance (AMED) and Hausdorff Distance (HD) (Song et al., 2010) defined as:

$$AD(A,B) = \frac{1}{2}\left[\frac{1}{m}\sum_{i=1}^{m} d(a_i,B) + \frac{1}{n}\sum_{j=1}^{n} d(b_j,A)\right]$$

$$AMED(A,B) = max\left[\frac{1}{m}\sum_{i=1}^{m} d(a_i,B), \frac{1}{n}\sum_{j=1}^{n} d(b_j,A)\right]$$

$$HD(A,B) = max\left[\max_{i\in 1,\ldots,m} d(a_i,B), \max_{j\in 1,\ldots,n} d(b_j,A)\right]$$

where $A = \{a_1, a_2, \ldots, a_m\}$ and $B = \{b_1, b_2, \ldots, b_n\}$ are the two contours to be compared and $d(a_i,B) = \min_{j\in 1,\ldots,n} ||a_i - b_j||$ is the distance from a_i to the closest point on contour B. These last measures are normalized by the ROI diagonal.

Note that the selected region based measures are measures of accuracy (the higher the better), while contour based measures are measures of error (the lower the better).

Classification tasks are evaluated using the Mean Absolute Error (MAE). When a more detailed analysis of the results is needed, True Positive (TP), True negative (TN), False Negative (FN) and False Positive (FP) rates are presented:

$$TPr = \text{Sensitivity} = \frac{TP}{TP+FN} \qquad FPr = \frac{FP}{TN+FP}$$

$$FNr = \frac{FN}{TP+FN} \qquad TNr = \text{Specificity} = \frac{TN}{TN+FP}$$

Evaluation of detection techniques is not straight-forward and the adopted evaluation technique will be presented in Chapter 5.

For the tasks that include a training phase, the dataset is divided into two non-overlapping sets: 75% of the data is randomly selected for training and the remaining 25% is used for testing. This splitting of the data is repeated 40 times in order to obtain more stable results for performance estimation. Special care was taken when dividing the dataset into train/test. For each one of the 40 splits, images from the same patient are forced to belong to the same set. Thus, it was not possible, for an algorithm to be trained on a CC image and tested on a MLO image of the same breast. This reduces model over-fitting and prevents results being over-optimistic.

All features are normalized within the train set to have zero mean and unit variance. Each model parametrization is selected by two-fold cross-validation inside the training set. Unless stated otherwise, the metric used in the two-fold cross-validation to select the parameters is the classification error. As suggested in (Hsu et al., 2003), exponentially growing sequences were used (for example, for SVMs - Support Vector Machines - with a Radial Basis Function - RBF - kernel, $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$). For k-Nearest Neighbours (kNN), a k value between 1 and 100 was selected, also by using two-fold cross-validation inside the training set.

Two results are considered as being "significantly different" if the difference is statistically significant at the 5% level according to a paired two sided t-test, where each pair of data points consists of the estimates obtained in one of the 40 runs of the two learning schemes being compared. In case a paired test is not possible, an unpaired test is performed and its use is explicitly mentioned.

2.9 Conclusions

The requisite background needed for understanding this dissertation was compiled in this chapter. Some notions on breast cancer and the most used imaging methods to assess this pathology were first described. Next, available databases were discussed and the motivation for creating a new database was presented. The new database, INbreast, will be used throughout this thesis to evaluate all the proposed methodologies. To assess the value of the proposed techniques it is important to know what is already available for the user. A compilation of commercial CAD systems for breast cancer along with their published performances was also presented in this chapter. Finally, the evaluation techniques used through this thesis were specified.

Chapter 3

Mammogram pre-processing

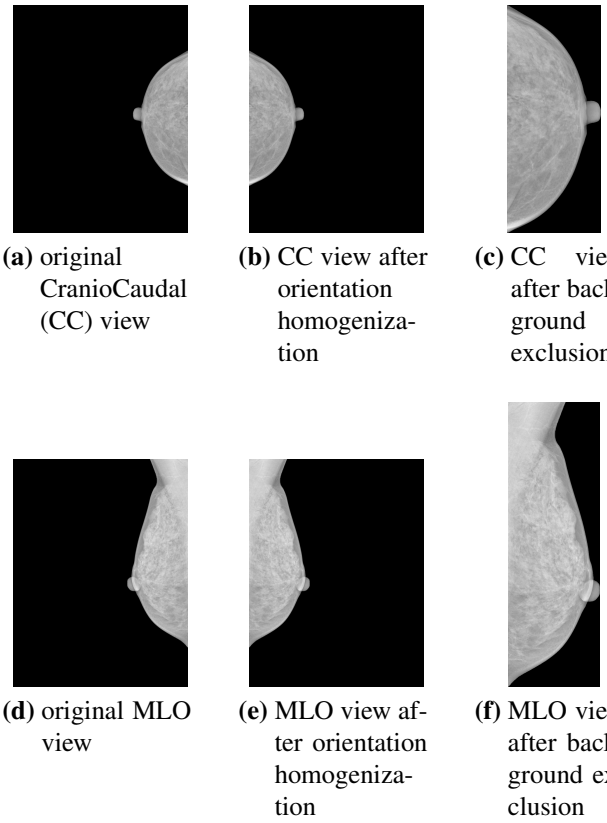
Pre-processing is the first step of every traditional image processing pipeline. In the case of mammograms, typical pre-processing techniques include: noise reduction (Romualdo et al., 2013), image enhancement (Wang et al., 2013), background exclusion (Li et al., 2013), orientation homogenization (Li et al., 2013), and pectoral muscle identification (Akram et al., 2013) among others.

In this work, it was considered that the Full Field Digital Mammography (FFDM) images available nowadays have sufficient quality that noise reduction and image enhancement techniques are not needed. Concerning background exclusion and orientation homogenization, standard techniques were used. Namely, the orientation of the mammograms is judged (by comparing the average intensity level of the right half with the left half) and those images where the nipple faces left are vertically mirrored. The foreground (pectoral muscle and breast area) is selected using Otsu (1979) threshold and the image is cropped to the bounding box around the breast (see Figure 3.1).

While orientation homogenization and background exclusion are trivial tasks, the pectoral muscle is not always easy to identify. Thus, more attention was given to this task. While not present in the CC view, the pectoral muscle represents a predominant density region in most MLO views of mammograms, and can affect the results of image processing methods. Intensity-based methods, for example, can present poor performance when applied to differentiate dense structures such as the fibro-glandular disc or small suspicious masses, since the pectoral muscle appears at approximately the same density as the dense tissues of interest in the image. The inclusion of the pectoral muscle in the image data being processed can bias the detection procedures, since some masses may have a density similar to the pectoral muscle density. Another important need to identify the pectoral muscle lies in the possibility that the local information of its edge, along with an internal analysis of its region, may be used to identify the presence of abnormal axillary lymph nodes, which may be the only manifestation of an occult breast carcinoma (Ferrari et al., 2004).

In this chapter[1], two proposed methods are described for pectoral muscle segmentation: Polar

[1] Publications related with this chapter include (Gaspar et al., 2011; Cardoso et al., 2010; Domingues et al., 2010).



(a) original CranioCaudal (CC) view

(b) CC view after orientation homogenization

(c) CC view after background exclusion

(d) original MLO view

(e) MLO view after orientation homogenization

(f) MLO view after background exclusion

Figure 3.1: Orientation homogenization and background exclusion examples. For better visual clarity, the scale was not kept constant between images.

coordinates and the shortest path (Section 3.2) and Shortest path with endpoints learnt by SVMs (Section 3.3). Before presenting the methods, a brief literature review on pectoral muscle detection is given (Section 3.1). The chapter ends with Section 3.5 where some conclusive remarks are presented.

3.1 Related work

Ganesan et al. (2013b) divided pectoral muscle segmentation methods into five groups: (1) intensity based approaches, (2) line detection techniques, (3) statistical methods, (4) wavelet based segmentation, and (5) other. In the next paragraphs one example of each group is given.

Li et al. (2013) employed two anatomical features of the pectoral muscle, homogeneous texture and high intensity deviation to identify the initial pectoral muscle edge. Then, a Kalman filter was used to refine the initial ragged edge. A qualitative evaluation was performed and the acceptable rate (contours subjectively classified as either "Exact" or "Optimal") was 90% and 92% for the mini-MIAS database and the DDSM database, respectively.

Kwok et al. (2004)'s method is a line detection technique. They proposed an adaptive algorithm that uses knowledge about the position and shape of the pectoral muscle. The pectoral

edge is first estimated by a straight line which is validated for correctness of location and orientation. This estimate is then refined using an iterative "cliff detection" to delineate the pectoral margin more accurately. The algorithm was applied to the entire MIAS database of 322 images. The segmentation results were evaluated by two expert radiologists, who rated 84% of the curve segmentations to be adequate or better.

Liu et al. (2014) presented an algorithm that uses the Anderson–Darling goodness-of-fit test to extract a feature image by assuming non-Gaussianity for the pectoral muscle boundaries. The method was applied on MLO view mammograms from the mini-MIAS database. The average False Positive (FP) and False Negative (FN) rates were 2.32% and 3.81%, respectively. The Average Distance (AD) was 2.12 ± 1.83 mm, and the Hausdorff distance was 3.27 ± 4.57 mm. The method cannot, however, determine if a result is the true pectoral muscle edge when a mammogram has no pectoral muscle or multiple pectoral muscle edges (e.g., folded pectoral muscle). The methods proposed here (Cardoso et al., 2010; Domingues et al., 2010), were also classified under the category of statistical methods.

The method of Ferrari et al. (2004) is an example of a wavelet based segmentation technique. They presented a multi-resolution technique using Gabor wavelets. The method starts by convolving a group of Gabor filters, specially designed for enhancing the pectoral muscle edge, with the region of interest containing the pectoral muscle. After computing the magnitude and phase images using a vector-summation procedure, the magnitude value of each pixel is propagated in the direction of the phase. The resulting image is then used to detect the relevant edges. Finally, a post-processing stage is used to find the pectoral muscle edge. The method was applied to 84 MLO mammograms from mini-MIAS. The average FPr and FNr were, respectively, 0.58% and 5.77% and the Hausdorff distances had a mean and *std* of 3.84 ± 1.73 mm.

Active contours were included in the "other" category. The work in (Akram et al., 2013) uses active contours combined with a stopping algorithm to obtain a contour which contains the boundary of the pectoral muscle. Accuracy values are only provided for 8 mini-MIAS images with an average of 87.5%.

As can be seen, most of the methods were tested on mini-MIAS, that has several shortcomings, for instance, the fact that it is a digitized and not a digital database, therefore it is not possible to have extremely high confidence in the spatial precision due to the digitization process[1]. Moreover, contour evaluation metrics are also not consistent over published papers and thus a fair comparison between methods is not possible.

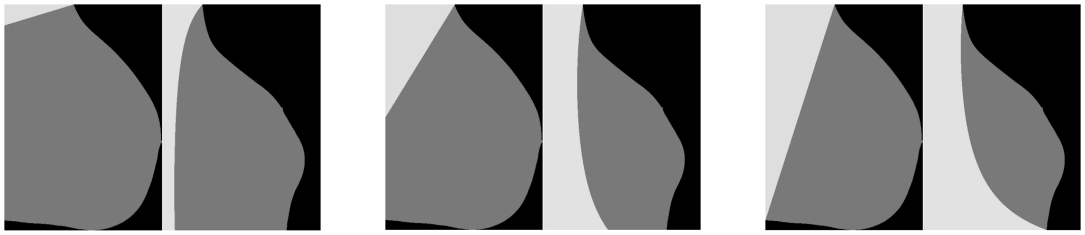
As part of the proposed complete breast analysis system in this thesis, two approaches for pectoral muscle detection are now presented. The first uses polar coordinates (Section 3.2), while the second uses SVMs to learn the pectoral muscle end-point positions (Section 3.3). They both make use of the shortest path algorithm (Gallo and Pallottino, 1988).

[1] A discussion on the available databases can be found in Section 2.5.

3.2 Polar coordinates and the shortest path (SPPC)

Intuitively, the muscle boundary manifests itself as a change in the grey-level values of the pixels, giving rise to an edge in the resulting image. Therefore, it can be argued that the muscle boundary corresponds to a path through edge pixels. If paths through edge pixels are favoured with the appropriate weight in the graph, the muscle boundary is a short path between the two endpoints. Efficient algorithms are available to solve this problem, such as the well-known Dijkstra (1959) algorithm.

A difficulty with searching for the shortest paths (shortest in the sense on minimizing the cost of the path) between the top row and left column is that small paths, near the top-left corner are naturally favoured. To overcome this challenge, the image is processed using polar coordinates, as in Figure 3.2. It is assumed that the origin of the coordinates is in the top-left corner. Note that the left column is mapped to the bottom row (corresponding to an angle of $\pi/2$ radians) and the top row stays in the same position. In this new coordinate-system, the path to search is now between the top and bottom rows.



(a) original image on the left and transformed image on the right

(b) original image on the left and transformed image on the right

(c) original image on the left and transformed image on the right

Figure 3.2: SPPC pre-processing: polar transformation toy examples. In the transformed images, the horizontal axis represents the radius and the vertical axis the angle, from 0 to $\pi/2$ radians.

In the transformed image, the simplifying assumption can be made that the vertical paths do not zigzag back and forth or up and down. Therefore, the search may be restricted among connected paths containing one, and only one, pixel in each row between the two end-rows.

The key steps involved in the shortest path computation encompass:

- a gradient computation of the original image. In a broader view, this can be replaced by any feature extraction process that emphasizes the pixels on the pectoral muscle boundary.

- to consider the gradient image as a weighted graph with pixels as nodes and edges connecting neighbouring pixels. To assign a weight w to an arc determined by the gradient values of the two incident pixels.

In this work, the weight of the arc connecting 4-neighbour pixels p and q is expressed as an exponential law (Oliveira, 2013, Ch.9):

$$\widehat{f}(g) = f_\ell + (f_h - f_\ell)\,\frac{\exp((255-g)\beta)-1}{\exp(255\beta)-1}, \tag{3.1}$$

with $f_\ell, f_h, \beta \in \mathbb{R}$ and g as the minimum of the gradient computed on the two incident pixels. For 8-neighbour pixels, the weight was set to $\sqrt{2}$ times that value. The parameters were experimentally selected as $f_\ell = 2$, $f_h = 32$ and $\beta = 0.025$.

The gradient model adopted in the experiments is based on the Prewitt (1970) operator. The Prewitt operator is applied on the x and y directions. From the computed values, G_x and G_y, the magnitude of the gradient is estimated as $z = \sqrt{G_x^2 + G_y^2}$.

3.2.1 SPPC algorithm

The SPPC algorithm can be implemented as a sequence of a few high-level operations, as presented in Figure 3.3.

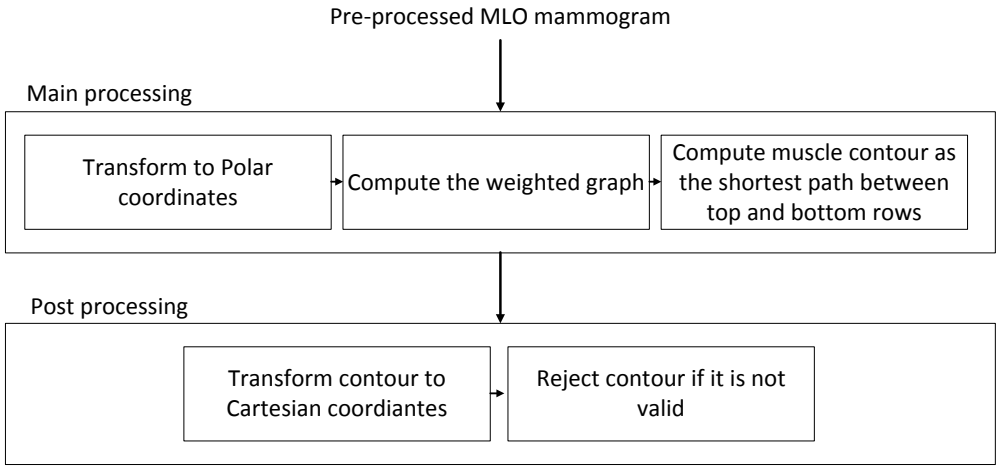
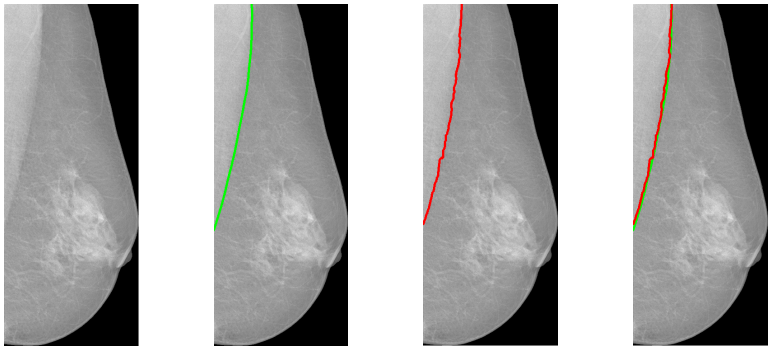


Figure 3.3: Main operations of the SPPC method.

Since the pectoral muscle may not be present in some MLO images and the main processing step of our approach always outputs a predicted boundary, it is necessary to validate if the detected path should be accepted or not as a muscle boundary. The validation rule, in the Cartesian coordinate representation, assumes that a muscle boundary runs "more or less monotonically" from the top margin to the left margin. In particular, it is not expected that the path deviates a lot to the right of the end-point on the top row; likewise, the path should also not go far below the end-point on the left margin. Therefore, if the column of rightmost point in the contour is more than k times the column of the end point on the top row or the row of the lowest point is more than k times the row of the end point on the left column, the contour is rejected and the mammogram is assumed

to be without a muscle boundary. In the experimental set-up reported in Section 3.4, k was set at 1.2 (acceptable values for this parameter range from 1.0 to 2.5).

Although a complete evaluation is presented in Section 3.4, an illustrative example is included in Figure 3.4.



(a) original image **(b)** GT contour **(c)** SPPC contour **(d)** superimposed GT and SPPC contours

Figure 3.4: SPPC example. Note that the contour has been made in polar coordinates and then mapped back to the original cartesian coordinate system.

3.3 Shortest path with endpoints learnt by SVMs (SPLE)

The main drawback of the SPPC method is the need to transform the image to the polar domain (and then back to the Cartesian domain). This can have as a consequence not only a larger computational burden, but also a possible loss in the image resolution. Another method was thus proposed, Shortest Path with Learnt Endpoints (SPLE). This approach has two main parts. As a first step, the endpoints of the muscle boundary are detected with the help of a regression model. Next, the muscle boundary is computed as the shortest path between the now known endpoints.

3.3.1 Automatic detection of the endpoints

The detection of the endpoints is based on techniques of supervised machine learning, namely regression models. The position of the end-point on the top row (ET) is normalized by the width of the image (after cropping). Likewise, the position of the end-point on the left column (EL) is normalized by the height of the image. Two Support Vector Regression (SVR) models (Schölkopf et al., 2000) were developed simultaneously. The first SVR model is trained to predict the ET, while the second SVR is trained to predict the EL.

The input features chosen to develop the SVR models are the grey-levels values obtained from a 32×32 thumbnail of the cropped mammogram. The SVR model predicting the ET is trained with data from the top half of the thumbnail. The SVR model predicting the EL is trained with

data from the left half of the thumbnail. An illustration is given in Figure 3.5. Therefore, the dimension of the feature data is 512 for each model.

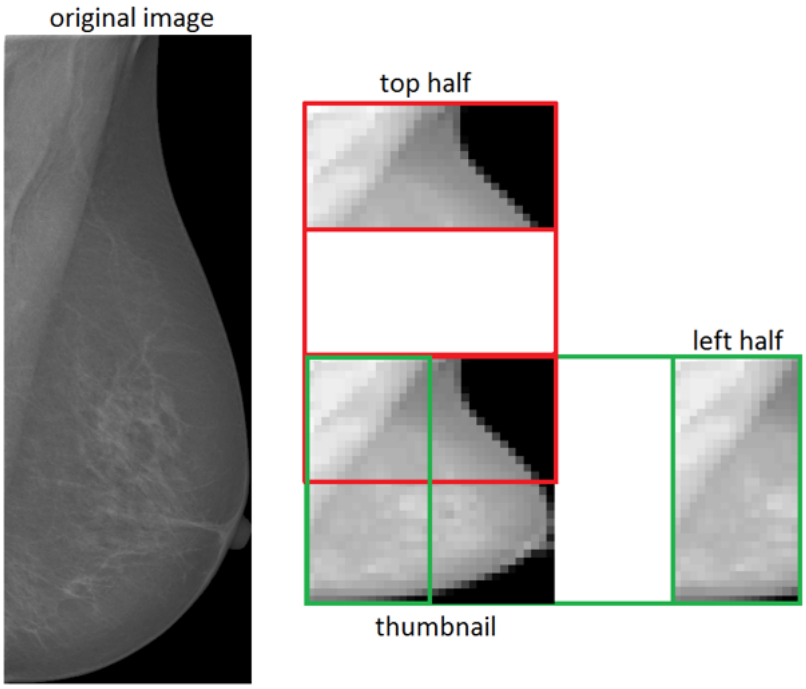


Figure 3.5: Illustration of the feature creation process for the SVR models.

3.3.2 Pectoral muscle boundary detection with known endpoints

Receiving the endpoints predicted by the SVR models as input, one is left with the computation of the muscle contour between the endpoints. As before, this problem is addressed by searching for the shortest path between the endpoints, after defining the weighted graph as described in Section 3.2.

3.3.3 SPLE algorithm

The SPLE algorithm can be implemented as a sequence of a few high-level operations, as presented in Figure 3.6. If the endpoints predicted by the SVR models are outside the valid range, the mammogram is assumed to be without muscle and the shortest path algorithm is not run.

Again, although a complete evaluation is presented in Section 3.4, an illustrative example is included in Figure 3.7.

3.4 Results

The tasks of orientation homogenization and background exclusion are trivial. They are thus not evaluated. Concerning the pectoral muscle segmentation, all MLO images from the INbreast were

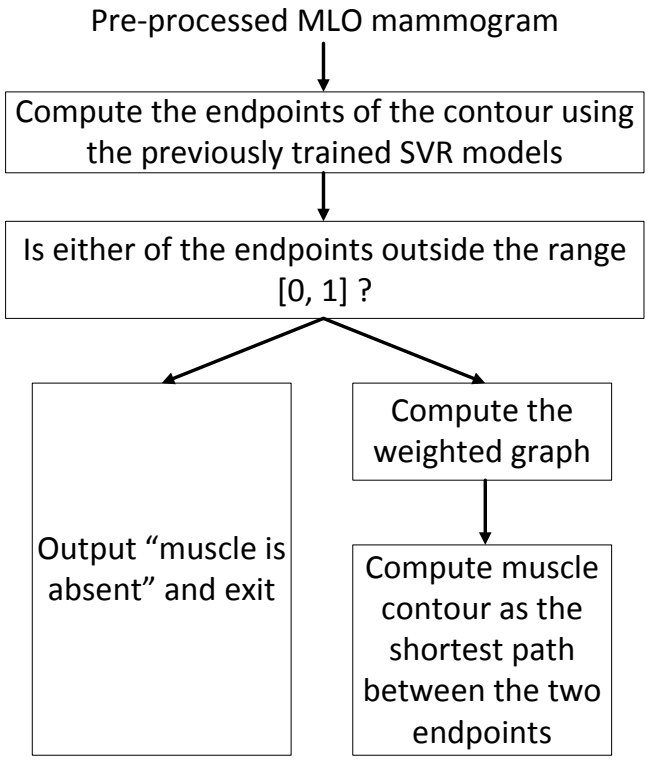
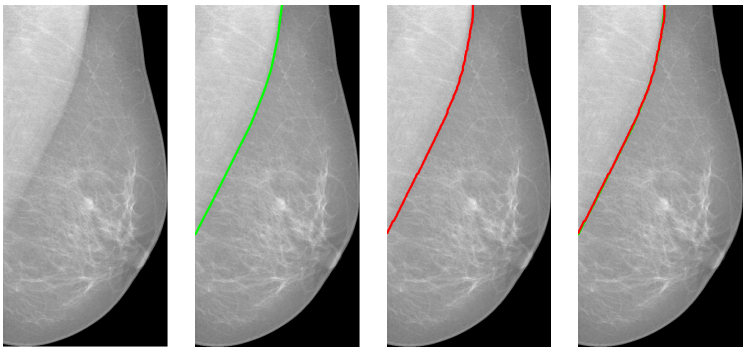


Figure 3.6: Main operations of the SPLE method.



(a) original image **(b)** GT contour **(c)** SPLE contour **(d)** superimposed GT and SPLE contours

Figure 3.7: SPLE example.

used. This sums to 206 mammograms. Pectoral muscle contours provided with INbreast were used as GT. For evaluation purposes, in the four images (out of 206) for which no contour is present in the mammogram, the reference contains only a single point at position $(0,0)$.

For the SPPC method, independently of the original size, the image in polar coordinates was

kept at a constant 1024×1024 size. This improves the computational performance without degrading the quality of the final result.

Results are presented in Table 3.1. Besides the proposed methodologies, a baseline segmentation method is also included. This baseline method simply connects a straight line from the mean GT start position (from all images in the database) to the mean end position. Another natural method to test is the Region Growing (RG) technique (Pham et al., 2000). It applies naturally to the pectoral muscle case, by using as seed the top left corner of the image[1]. The difficulty with RG method is to choose an appropriate threshold value. In the experiments, the best value in the set $[0.05\ 0.10\ 0.15\ 0.20\ 0.25]$ was used for each image.

Table 3.1: SPPC method: overall results in the position of the muscle boundary. Results are presented in the format: mean (*std*). Remember that AOM and CM are measures of accuracy (the higher the better), while AD, AMED and HD are measures of error (the lower the better).

Method	AOM	CM	AD	AMED	HD
Baseline	0.552 (0.219)	0.698 (0.146)	0.048 (0.037)	0.052 (0.039)	0.119 (0.078)
RG	0.216 (0.254)	0.471 (0.172)	0.193 (0.121)	0.209 (0.122)	0.391 (0.193)
SPPC	**0.658** (**0.325**)	**0.769** (**0.220**)	0.055 (0.083)	0.064 (0.089)	0.158 (0.175)

Region Growing is significantly worse than both the baseline and SPPC and for every studied metric. When comparing the Baseline and SPPC, SPPC is significantly better for both region metrics, but significantly worse when considering the HD metric. Figure 3.8 shows some of the results. A propensity to miss the part of the contour next to the left margin, where the boundary is usually more diffuse, was observed in the SPPC method. Moreover, when there are multiple strong edges, the algorithm may pick the wrong one.

In the SPLE method, the SVR model was based on the SVM package LIBSVM (Chang and Lin, 2011), implemented with the epsilon-SVR and a linear kernel. The average mean square error over the 40 repetitions for the ET position is 0.186 ± 0.010 pixels and for the EL is 0.415 ± 0.024 pixels. A trend was observed for the predicted ET to be on the left, and the predicted EL to be above the GT points. Next, the quality of the shortest path to find the muscle contour with the predicted endpoints was evaluated. Table 3.2 summarizes the results. A comparison with SPPC, the Baseline method and the Region Growing is included. Note that for these methods the values do not match exactly the ones in Table 3.1 due to the 40 different splits of the data into train/test needed to create the SVR models for SPLE. To understand if the main source of errors was from the end-point prediction or from the shortest path method, the shortest path method was re-computed using the true endpoints obtained from the reference contours as input. These results are presented in the last row of Table 3.2.

Once again, Region Growing is significantly worse than all the other methods and for every metric. When comparing Baseline with either SPPC or SPLE, differences are only significant for the more sensitive metric HD. The differences between SPPC and SPLE are not significant for any metric. SPLE with true endpoints is significantly better than all the other methods and for every metric. Examples for the full SPLE method are also included in Figure 3.8.

[1] Code available at
`http://www.mathworks.com/matlabcentral/fileexchange/32532-region-growing-2d3d-grayscale`.

Table 3.2: SPLE method: overall results in the position of the muscle boundary. Results are presented in the format: mean (*std*). Remember that AOM and CM are measures of accuracy (the higher the better), while AD, AMED and HD are measures of error (the lower the better).

Method	AOM	CM	AD	AMED	HD
Baseline	0.559 (0.030)	0.702 (0.020)	0.048 (0.004)	0.051 (0.005)	0.117 (0.009)
RG	0.211 (0.029)	0.468 (0.020)	0.193 (0.015)	0.210 (0.015)	0.393 (0.026)
SPPC	0.662 (0.051)	0.772 (0.035)	0.055 (0.012)	0.063 (0.013)	0.159 (0.026)
SPLE with true endpoints	0.927 (0.014)	0.949 (0.011)	0.005 (0.002)	0.005 (0.002)	0.015 (0.004)
SPLE	0.597 (0.035)	0.722 (0.024)	0.056 (0.009)	0.061 (0.010)	0.153 (0.018)

3.5 Conclusions

Image manipulation can have a strong impact on the success of subsequent tasks. A typical preprocessing applied to mammogram images is the removal of the pectoral muscle region. In this chapter, two methods for segmentation of the pectoral muscle are presented, namely polar coordinates and the shortest path (SPPC) and shortest path with endpoints learnt by SVMs (SPLE). Differences between SPPC and SPLE are not significant, however, the results for the SPLE method indicate that if a robust estimation of the endpoints can be achieved, then the pectoral muscle boundary can be effectively predicted using the shortest path. In fact, from the results, the prediction of the endpoints seems to be the main source of errors for the global estimation of the contour.

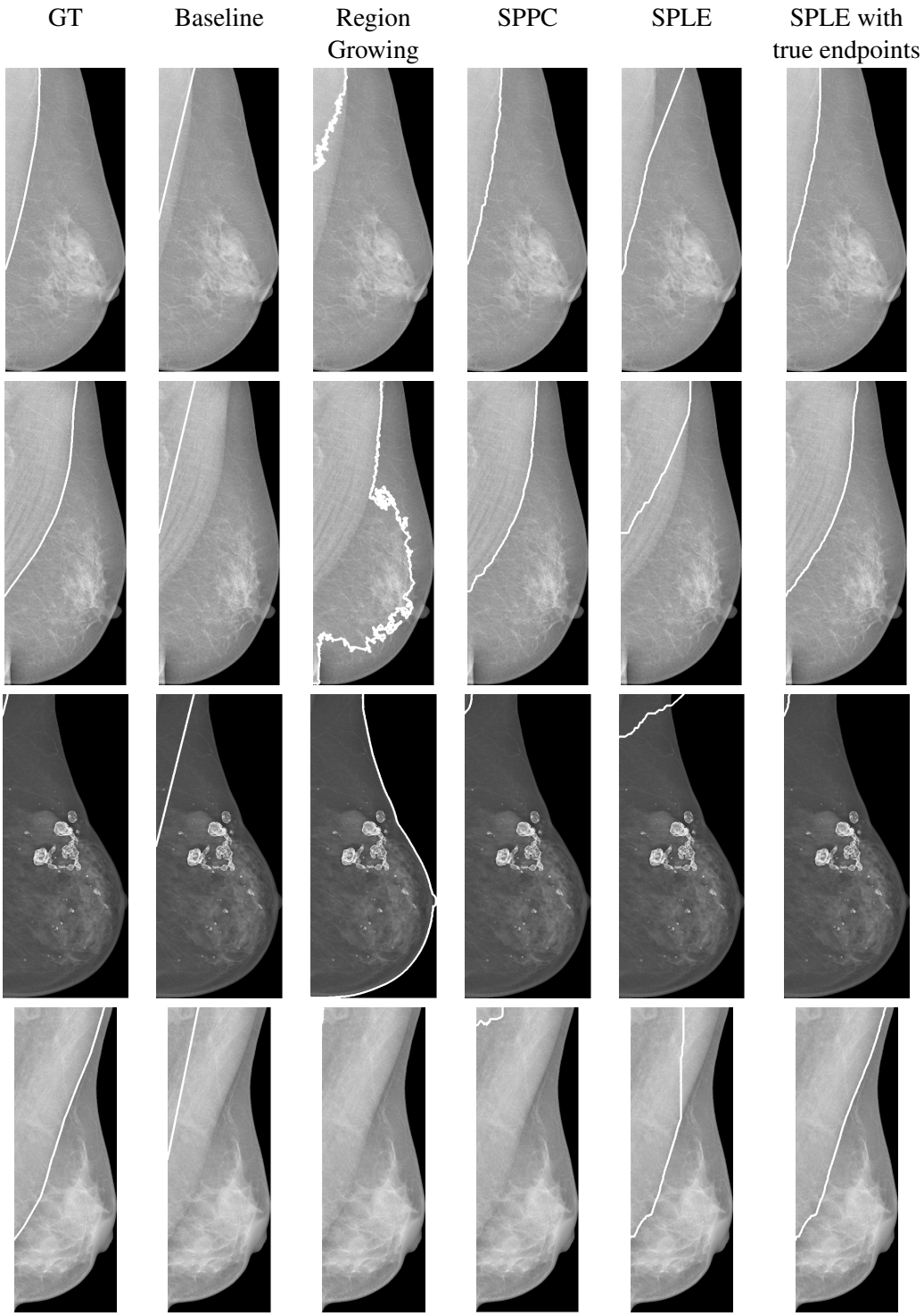


Figure 3.8: Pectoral muscle segmentation examples ordered per row as follows: GT, Baseline, Region Growing, SPPC, SPLE and SPLE with true endpoints

Chapter 4

Breast cancer screening

Breast cancer screening can be a very tedious and tiring task for the specialist. In general, only 0.5% of mammograms present an anomaly (Sun et al., 2004b; Elshinawy et al., 2010b). The assumption made here is that easy normal cases can be automatically detected, alleviating the human effort and giving the specialist more time to carefully evaluate more ambiguous cases. Although several Computer-Aided Detection and Diagnosis (CAD) systems have been developed and studied, little research has been done on the development of Screening systems (also known as Pre-CADs or Normal mammogram detection or analysis).

This Chapter starts with a description of the breast cancer screening program currently implemented in Portugal (Section 4.1). The state of the art in existing automatic screening systems is summarized in Section 4.2. The proposed approach to deal with this problem is detailed in Section 4.3[1]. The breast images are first automatically divided into two sets: dense and non-dense breasts. Then, two classifiers are developed, one specific to dense images and other specific to non-dense images that filter the suspicious mammograms. Results are presented in Section 4.4 and Conclusions in Section 4.5.

4.1 Breast cancer screening in Portugal

The Portuguese Breast Cancer Screening Program is a service regionally organized in partnership between "Liga Portuguesa Contra o Cancro" and "Administração Regional de Saúde". The Program is integrated in the European structure of quality control, through the Europe Against Cancer Program.

The flow of information of the Portuguese Breast Cancer Screening Program is as shown in Figure 4.1. The Program has three main phases:

- Convocation
  - demographic data of all women aged between 45 and 69 is requested from the health centres;

[1] Publications related with this chapter include (Domingues et al., 2012b; Fonseca, 2013; Zolfagharnasab et al., 2013; Bessa et al., 2014).

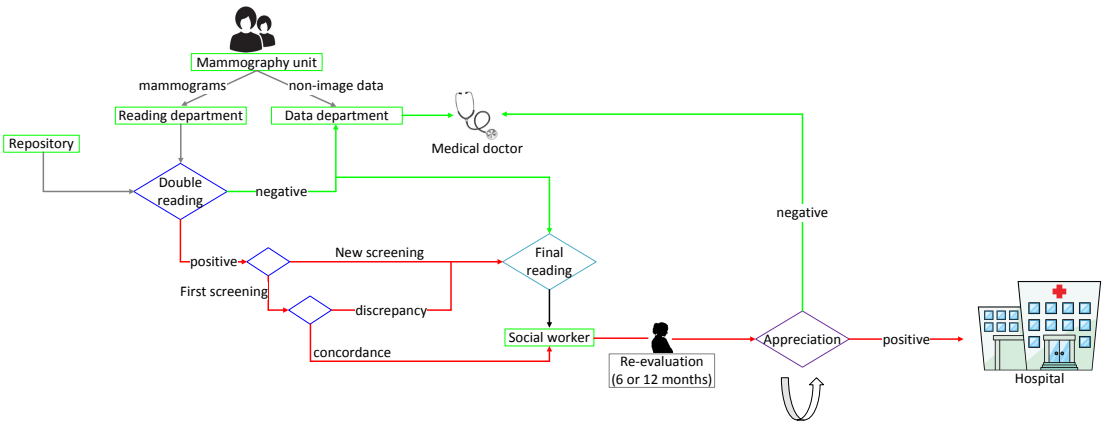


Figure 4.1: Flow of information of the Portuguese breast cancer screening program.

  - automated comparison with the list of users already integrated in the Breast Cancer Screening Program;
  - personal and customized call to the target population, specifying the place, date and time of the screening.

- Implementation
  - reception of the women in the mammography unit;
  - collection of personal, demographic and clinical data;
  - mammography execution;
  - personal information and images are sent to the coordination centre;
  - double-blind-reading and consensus reading in case of discrepancy:
    * In case of a **negative** result, the information is forwarded to the user and the Family Doctor and a subsequent notice is made within two years;
    * In case of a **positive** result, a personal contact is made with the woman and she is directed to a specialized appointment in a dedicated centre.
  - In the specialized appointment the patients are classified as having a negative, a questionable, or a positive result:
    * In case of a **negative** result, the information is forwarded to the user and the Family Doctor and a subsequent notice is made within two years;
    * In case of a **positive** result, the information is sent to the reference hospital, with instant scheduling of the date and time of the hospital appointment;
    * In case of a **questionable** result, a new specialized appointment is scheduled in a time frame of 6 or 12 months.

- Monitoring and evaluation

– six months after the date of the first entry in the reference hospital, the Program requests the hospital data that will be used to monitor and evaluate the Program.

A possible way to introduce an automatic screening system in the current practice is to replace one of the radiologists during the double-reading by the automatic decision. In the case that the automatic decision disagrees with the decision of the radiologist, the exam would be sent to the consensus reading.

4.2 Related work

Sun et al. (2004b,a), following (Sun, 2004)'s doctoral work, propose a two-stage system for normal mammogram identification. After a pre-processing stage, four types of features are extracted from overlapping blocks of size 512×512, namely curvilinear features, texture features, Gabor features and multi-resolution features. Each block is then classified by a Constrained Binary Decision Tree. If the result is not abnormal, the block is classified as Normal. In case the block is abnormal, a second classifier, this time a linear classifier, is used to further distinguish between Suspicious and Normal blocks. As there is an overlap between blocks, each pixel is classified five times. A majority voting is used to make the final pixel decision. Finally, the full image is considered as suspicious if one or more pixels are abnormal, otherwise, the mammogram is classified as normal. Experiments made on Digital Database for Screening Mammography (DDSM) retrieved a True Positive rate (TPr) of 0.82 and a True Negative rate (TNr) of 0.75. It was noted by the authors that most of the misclassifications were due to high breast density of the mammograms.

A Support Vector Machine (SVM) based method was introduced in (Chiracharit et al., 2007). The same pre-processing and features as in (Sun et al., 2004b,a) were used. Crossed-distribution feature pairs were identified and mapped into new features that could be separated by a zero-hyperplane of the new axis. The probability density functions of the features of normal and abnormal mammograms were then sampled and the local probability difference functions are estimated to enhance the features. From a balanced set of 1 000 mammograms from DDSM, half were used for training and the remaining half for testing. The classification results show performances with 90% sensitivity and 89% specificity.

A group with strong contributions in this field is led by Elshinawy (2010). They noticed that most misclassification of abnormal tumours happen in dense breasts. They have thus used the density information to separate mammograms into fatty and dense. In (Elshinawy et al., 2010b), a traditional two-class classifier is compared with a one-class classifier. For the one class classifier, only the normal features are used for training while in the test phase, both normal and abnormal features are included. The assumption behind the use of the one class classifier approach is that the normal mammograms have less variability when compared to the abnormal mammograms which have a very wide variation of lesion type and shape. Classification was made per block and a majority voting approach was used to classify the full breast by combining CranioCaudal (CC) and MedioLateral Oblique (MLO) information. Results on the DDSM database show that separating

the mammograms according to density reduces the False Negative rate (FNr) while keeping the False Positive rate (FPr) as low as possible. Moreover, the one-class SVM outperformed the two-class SVM. After selecting the appropriate classifier, the team studied the feature extraction step. In (Elshinawy et al., 2010a), three different sets of textural features based on Local Binary Patterns (LBP) were assessed, namely, (1) simple LBP, (2) multi-resolution grey-scale and rotation invariant LBP, and (3) LBP based on the Fourier transform. LBP has two major advantages, it is invariant against monotonic grey level changes and it is computationally simple. Results on DDSM showed that LBP based on the Fourier transform outperformed the other two LBP-based methods. Grey Level Co-Occurrence Matrix (GLCM) features were studied in (Elshinawy et al., 2011b). Contrast, homogeneity, correlation, energy and entropy were obtained from four GLCM matrices at distance $d = 1$ and directions 0°, 45°, 90°, and 135°. The performance of using the mean of the obtained parameters was compared with the performance when using not only the mean but also the standard deviation (*std*). Results show that first and second order statistics, when used together, outperform using only first order statistics. Finally, in (Elshinawy et al., 2011a), LBP and GLCM features were compared. It was observed using DDSM that GLCM is a good feature for fatty-tissue, while LBP behaves better in dense-tissue mammograms. Moreover, the sensitivity for each tissue type was improved when compared to the sensitivity of using all mammograms regardless of tissue type. Results of the recall rate varied from 97.34% to 99.75%.

The work by Fonseca et al. (2013) and Fonseca (2013) follows the above approach of separating the breast accordingly to tissue type. There, GLCM and LBP features were studied in combination with several classifiers: k-Nearest Neighbours (kNN), SVM and Random Forests. Results using the 203 CC images from INbreast disagree with the initial hypothesis and point to the possibility that the initial separation of the images is not essential. It was also concluded that multi-resolution LBP features with a SVM classifier led to the best results, obtaining an accuracy of 95.35% with a TPr of 100% and a TNr of 90.69%.

The paper from Ganesan et al. (2014) presents a one-class classification pipeline for the classification of mammogram images into normal and abnormal. Because of the sparse distribution of abnormal mammograms, the two-class classification problem is reduced to a one-class outlier identification problem. The trace transform, which is a generalization of the Radon transform, is applied to extract the features. Using a private database, an accuracy rate of 92.48% was obtained with a Gaussian mixture model classifier.

In the study of Zhang and Zhang (2014), three combinations of wavelet and Fourier features, including Db2, Db4, and Bior 6.8 were tested. Classification into normal or suspicious using a private database was made with three classifiers, including a Back-propagation Network, Linear Discriminant Analysis (LDA), and Naive Bayes. Best results were attained using Bior and Fourier Features, where LDA achieved: TP= 301, FP= 19, FN= 220, TN= 130 Sensitivity= 57.8%, Specificity= 87.2%, Accuracy= 64.3%; Back-propagation Network: TP= 500, FP= 21, FN= 21, TN= 128, Sensitivity= 95.9%, Specificity= 85.9%, Accuracy= 93.7%; and Naive Bayes: TP= 510, FP= 68, FN= 11, TN= 81, Sensitivity= 97.8%, Specificity= 54.4%, Accuracy= 88.2%.

Tan et al. (2014b) analysed global mammographic texture and density features computed from

four-view images. First, global texture features related to the mammographic density distribution were extracted. Second, the computed features were given to two ANN classifiers that were separately trained on CC and MLO view images, respectively. Finally, the two ANN classification scores were combined using an adaptive scoring fusion method that automatically determined the optimal weights to assign to each view. A private database of FFDM images acquired from 1 052 recalled women (669 positive for cancer and 383 benign) was used to assess the results. An area under the Receiver Operating Characteristic (ROC) curve of 0.793 ± 0.026 was obtained.

Table 4.1 compiles the main features used in the above mentioned works. These features can either be extracted from the full breast region or from blocks belonging to the breast region and the results then merged, usually with a majority voting approach.

Table 4.1: Summary of features used in screening works.

Type	Features
Curvilinear	Line pixel count, diagonal half line pixel counts, half ratios, means, *std*s and entropy of angles, local lines and local angles (Sun et al., 2004a,b; Chiracharit et al., 2007)
Fourier	Mean, *std*, skewness and kurtosis (Zhang and Zhang, 2014)
Wavelet	Mean, variance, compactness, fractal dimension and entropy of each sub-band (Sun, 2004; Sun et al., 2004a,b; Chiracharit et al., 2007)
	Mean, *std*, skewness and kurtosis (Zhang and Zhang, 2014)
Gabor	16 filters (Elshinawy, 2010)
	Means and *std*s of Energy (Sun et al., 2004a,b; Chiracharit et al., 2007)
GLCM	Contrast, correlation, energy, homogeneity and entropy (Elshinawy et al., 2011b,a; Fonseca et al., 2013; Fonseca, 2013)
	Energy, entropy, their sums and differences, maximum probability, correlation, diagonal correlation, inertia, homogeneity, H_{xy}, sums of variance, shade and prominence (Sun et al., 2004a,b; Chiracharit et al., 2007)
LBP Fourier	Resolutions 8, 16 and 24 with radius 1, 2 and 3 (Elshinawy, 2010; Elshinawy et al., 2010a, 2011a)
LBP multi resolution	Mean, *std*, entropy, energy and skewness (Fonseca et al., 2013; Fonseca, 2013)
	Resolutions 8 and 16 with radius 1 and 2 (Elshinawy, 2010; Elshinawy et al., 2010a, 2011a)
LBP simple	Contrast (Elshinawy, 2010; Elshinawy et al., 2010a, 2011a)
	Mean, *std*, entropy, energy and skewness (Fonseca et al., 2013; Fonseca, 2013)
Trace transform	Diametric functional followed by circus function (triple feature) (Ganesan et al., 2014)

In the above summary, only works that explicitly classify the full mammogram (or breast) into suspicious or not suspicious were included. For related works but with slightly different strategies (using, for instance, only a ROI and not the full mammogram image), consult (Liu, 1999; Sun et al., 2002; Mini and Thomas, 2003; Petroudi and Brady, 2006; Eddaoudi et al., 2006; Mini, 2011; Deepak et al., 2012), among others.

The idea of splitting the mammograms according to the density is also adopted here. As the goal is to have a fully automatic system, the Ground Truth (GT) information will not be used, as done in the above mentioned works, but a classifier was created instead. Contrary to screening, there is a wide amount of research dedicated to the breast tissue classification problem. Thus, only some recent works will be referred to here. The works (Oliver et al., 2006) and (Oliver, 2008, Ch.3) present a deeper overview.

There are several ways to approach the density classification problem. In (Chen et al., 2013), for instance, the overall profile of breast tissue density was represented using a topographic map obtained from the upper level sets of an image. A shape tree was constructed to describe the

topological and geometrical structure of the shapes within the topographic map. Two properties, saliency and independence, were defined to detect shapes of interest based on the shape tree. A density map was further generated focusing on dense regions to provide a quantitative description of breast density. Finally, risk classification was performed based on the breast density measures derived from the density map. The validity of the method was evaluated using MIAS and a subset of 1 662 mammograms from DDSM. The obtained classification accuracies were 76.01% and 81.22% for MIAS and DDSM, respectively.

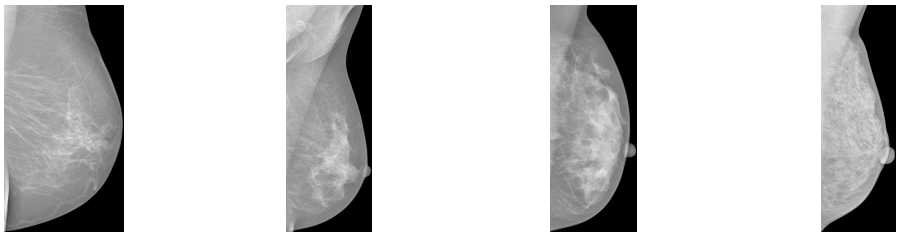
Recently, Llobet et al. (2014) presented a semi-automated and a fully automated tools to assess breast density. The first tool is based on a supervised interactive thresholding procedure. The automatic method relies on pixel-level labelling, allowing density measurement on a continuous scale. It combines a classification scheme based on local features and thresholding operations that improve the performance of the classifier. A private dataset of 655 mammograms is used. Three expert radiologists measure the density in each of the mammograms using the semi-automated tool. Density is also measured by the fully automated system and the correlation between both methods is computed. The results show an average Intraclass Correlation Coefficient (ICC) of 0.922 among raters when using the semi-automated tool, whilst the average correlation between the semi-automated and automated measures is ICC = 0.838.

Sharma and Singh (2014) propose a hybrid scheme for classification of fatty and dense mammograms using correlation-based feature selection (CFS) and sequential minimal optimization (SMO). Texture analysis is done on a region of interest selected from the mammogram. Various texture models are used to quantify the texture of parenchymal patterns of breast. To reduce the dimensionality and to identify the features which differentiate between breast tissue densities, CFS is used. The selected features are: correlation and inverse difference moment of GLCM; entropy of Grey level difference statistics; mean and skewness of first-order statistics; the Ripple Edge of Law's texture energy measure; the Hurst coefficient at resolution 2; and the angular sum of the Fourier power spectrum. Finally, classification is performed using SMO with a polynomial kernel of degree one. The performance is evaluated using 322 images of the MIAS database and an accuracy of 96.46% is obtained with a sensitivity of 100% and a specificity of 88.23%.

The work presented by Constantinou et al. (2014) investigates the use of Amplitude-Modulation Frequency-Modulation models in the evaluation of multi-scale Instantaneous Amplitude (IA) features for the characterization of breast density. Normalized histograms of the IA across the different frequency scales - estimated using multi-scale Dominant Component Analysis - are used to model the breast density classes. Classification of a new mammogram into one of the density categories is achieved using the k-nearest neighbour method with the Euclidean distance metric. The method, when evaluated using the MIAS database, had a breast density classification accuracy of more than 80%.

4.3 Normal breasts identification

Although it is expected that normal mammograms have less variability than abnormal ones, normal mammograms can still present very different appearances, as shown by the eight different normal mammograms in Figure 4.2.



(a) ACR I **(b)** ACR II **(c)** ACR III **(d)** ACR IV

Figure 4.2: ACR breast tissue types.

It can be seen from Figure 4.2 that part of the visual differences are due to breast density. Three main reasons have been identified for making a system dependent on breast density, namely:

- density has been associated with a higher risk of cancer (Chen et al., 2013);
- masses and calcifications are also dense and are thus harder to detect in dense breasts (Gils et al., 1998);
- density decreases the sensitivity of automatic systems (Tortajada et al., 2012).

The architecture used in this work for breast screening is depicted in Figure 4.3. Although other works, for example (Elshinawy et al., 2011a; Fonseca et al., 2013), have used the separation of the breasts according to density, to the best of our knowledge, this is the first work that includes an automatic separation in a screening context. The other works (Elshinawy et al., 2011a; Fonseca et al., 2013) have based the separation on GT information.

For the density classification, three different subsets of features were tested:

- intensity: the mammogram is re-sized to 32×32 and all the intensity values are used as features (total of 1024 features);
- statistics: breast intensity *std*, 75^{th} percentile (q_3), and number of pixels with intensity smaller than $q_1 - 1.5 \times (q_3 - q_1)$ with q_1 being the 25^{th} percentile (total of 3 features);
- histogram: a histogram of 32 bins is created for the breast intensity values (total of 32 features).

The motivation for the above features is that density is mainly reflected in a mammogram image by its intensity values. Considering the GT information, ACR breast density varies from I to IV (Table 2.2). In this work it was considered sufficient to distinguish only between two classes that will be referred to as non-dense (ACR I and ACR II) and dense (ACR III and ACR IV). A

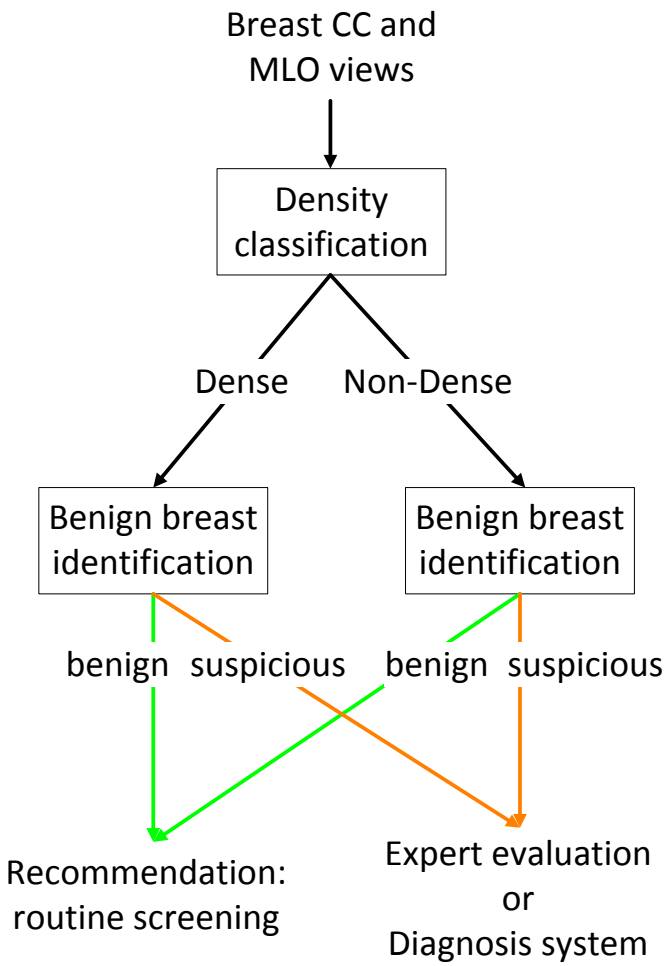


Figure 4.3: Screening block diagram.

more granular division could lead to several problems, in particular less data to train and test the models and consequently a less accurate classification. Experiments with different classifiers are presented in Section 4.4.

For the benign breasts identification blocks, seven different subsets of features were tested:

- Gabor: The Gabor transform provides simultaneous localization in both the spatial and frequency domain (Kovesi, 1999; Kuse et al., 2011). Features are extracted from the frequencies of the Gabor filter-bank (Chiracharit et al., 2007). Mean-squared energy and mean amplitude are computed for 3, 4, 5 and 6 wavelet scales and 4 and 6 filter orientations;

- $GLCM_1$: In this work, as in (Fonseca et al., 2013; Fonseca, 2013), to obtain the GLCM matrices, a distance of $d = 1$ and four directions, 0°, 45°, 90° and 135° were used. Five descriptors were extracted from each GLCM, namely, contrast, correlation, energy, homogeneity, and entropy. The final feature vector is the mean and *std* of each descriptor over the four GLCM matrices;

- $GLCM_2$: Another set of features based on the GLCM matrix was also tested. The same values for distance and directions were used, but this time 13 descriptors were extracted, namely, Entropy, Contrast, Correlation, Difference Entropy, Difference Variance, Energy, Information Measure of Correlation 1, Information Measure of Correlation 2, Inverse Difference Moment, Sum Average, Sum Entropy Sum of Variances, and Sum Variance (Haralick et al., 1973). The final feature vector is not the mean and *std* as above, but simply the concatenation of the descriptors over the four GLCM matrices;

- GLRL: Each element (r,c) in the GLRL matrix, represents the probability of occurrence of a pixel having grey level values r and run-length c in the image. Four directions were used (0°, 45°, 90° and 135°) and 11 parameters were extracted for each direction: Short Run Emphasis, Long Run Emphasis, Grey-Level Non-uniformity, Run Length Non-uniformity, Run Percentage, Low Grey-Level Run Emphasis, High Grey-Level Run Emphasis, Short Run Low Grey-Level Emphasis, Short Run High Grey-Level Emphasis, Long Run Low Grey-Level Emphasis, and Long Run High Grey-Level Emphasis (Tang, 1998);

- InvMom: These features correspond to the Seven Invariant Moments for the image (Hu, 1962);

- LBP: Five features (mean, *std*, entropy, energy and skewness) are extracted from the 256 bins histogram of the LBP image (Fonseca et al., 2013; Fonseca, 2013);

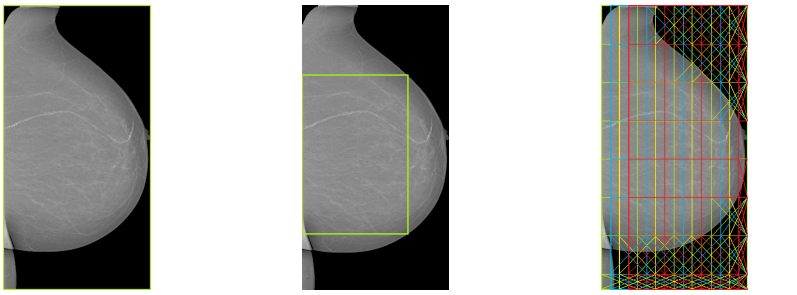
- Zernike: This vector of features corresponds to the Zernike Moments for orders from 3 to 17 with repetition (Tahmasbi et al., 2011).

As seen in Table 2.3, Breast Imaging Reporting And Data System (BI-RADS) class 1 corresponds to breasts with no findings. Binary classification was thus made with the goal of separating BI-RADS 1 from all the other BI-RADS classes.

Features were extracted in three different ways: (1) from the breast ROI, (2) from the largest (in the sense of area) rectangle (aligned with the image axis) inside the breast region, and (3) from overlapping squares with length of one quarter of the breast ROI width. In the last case, squares with fewer than 50% of pixels inside the breast were discarded and the mean and *std* of all squares for each feature was taken. Figure 4.4 illustrates each feature extraction method (for the overlapping squares technique, only horizontal translations are shown. Similar translations were also made on the vertical direction).

When selecting the model parametrization (for each benign breasts identification block), special care was taken in order to reduce the False Negative (FN) fraction. It is important that no suspicious image is classified as normal, otherwise, a possible cancer may be detected only in the next screening session (usually two years later) and the chances of cure decrease. More details on this topic are given in the experimental part of this work (Section 4.4).

If the final classification is benign, the general, country specific, routine screening is advised. However if suspicious, no automatic decision is made and the breast image is passed to a specialist (a diagnosis system can also give the specialist a second opinion on the malignancy).



(a) breast ROI **(b)** biggest rectangle inside the breast region **(c)** overlapping squares

Figure 4.4: Regions of analysis for feature extraction.

4.4 Results

The first part of this section (Subsection 4.4.1) presents the evaluation of the density classification block. The second part (Subsection 4.4.2) focus on the identification of normal breasts block(s).

4.4.1 Density classification

Three types of features were tested for the density block: (1) 1024 intensity values of the breast region; (2) 3 statistics extracted from the intensity values; and (3) an intensity histogram with 32 bins. In relation to the classifiers, kNN, LDA, and SVMs with two different kernels were used. kNNs were chosen due to their simplicity, LDA because it is a linear classifier and consequently it may have a more robust behaviour, and SVMs because they have been proved to be the state of the art classifier in several applications. Also, a majority voting ensemble of the 3 previous classifiers trained in each feature space was included. Results are shown in Table 4.2.

Table 4.2: Density classification results using every image in an independent way. Results are presented in the format: mean (*std*).

Features	Classifier	Error
intensity	kNN	0.316 (0.077)
	LDA	0.303 (0.088)
	SVM RBF	0.290 (0.063)
	SVM Sigmoid	0.276 (0.068)
statistics	kNN	**0.214** (0.058)
	LDA	0.235 (0.073)
	SVM RBF	**0.222** (0.067)
	SVM Sigmoid	0.305 (0.114)
histogram	kNN	0.245 (0.057)
	LDA	0.341 (0.064)
	SVM RBF	0.244 (0.064)
	SVM Sigmoid	0.309 (0.090)
ensemble	kNN	**0.220** (0.067)
	LDA	0.283 (0.089)
	SVM RBF	**0.215** (0.064)
	SVM Sigmoid	0.284 (0.087)

The kNN classifier with statistic features is significantly better than all the other tests with the exception of the SVM RBF with statistic features, the ensemble kNNs and the ensemble SVM RBF. As the ensemble options and SVM classifiers are computationally more demanding, the kNN classifier with statistic features will be used in the remaining experiments.

Results in Table 4.2 were obtained treating the 410 INbreast images independently. Next, in Table 4.3, results when using only the CC view, only the MLO view, and combining CC and MLO view (by concatenating the features extracted from the CC view with the features extracted from the corresponding MLO view) are shown.

Table 4.3: Density classification results per view. Results are presented in the format: mean (*std*).

view	Error
CC	0.222 (0.112)
MLO	0.236 (0.076)
both	**0.195** (0.074)

Differences are significant when comparing results using both views compared to the MLO view alone. However, the difference is not significant when comparing results using both views with either kNN with statistic features in Table 4.2 or when using the CC view only. In the remaining experiments, breast density will be classified with kNN and using as features the 3 statistics from the CC view concatenated with the 3 statistics from the MLO view. Some example results can be seen in Figure 4.5.

Note that several other features and combinations could be tested. However, with this block, the aim is to have a simple and fast classifier that separates the data into two meaningful sets. It is not necessary that it perfectly mimics the GT density classification.

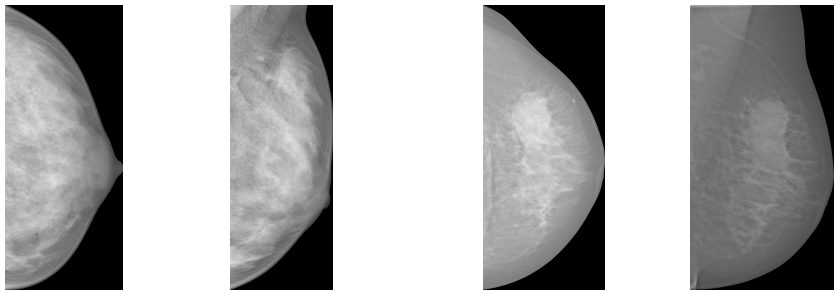
4.4.2 Normal breasts identification

As described in Section 4.3, features were extracted in three different ways: (1) from the breast ROI, (2) from the biggest rectangle inside the breast region, and (3) from overlapping squares with size length of one quarter of the breast ROI width.

Besides these combinations of features and classifiers, the tests were divided into three sets: (1) no aggregation of the breasts according to density; (2) using the GT information to aggregate the breasts according to density; and (3) aggregating the breasts with the classification method evaluated in Section 4.4.1.

When selecting the parameters by two-fold cross validation, the relation $1.5FN + FP$ was minimized. This is to avoid suspicious breast images being classified as benign (by more strongly weighting the false negative component). Some selected results are presented in Tables 4.4, 4.5 and 4.6.

When comparing Table 4.4 with Table 4.5, the results are seen to improve. Specifically, the differences between the lines marked with $*$ are significant for every metric. For the lines marked with $**$, differences are significant for the True Negative (TN) and FP metrics. This supports

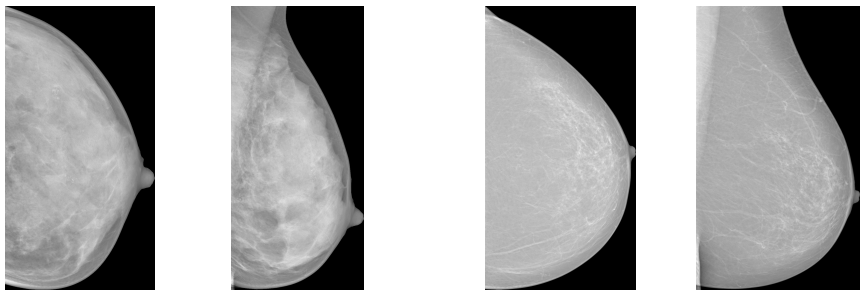


(a) correctly classified dense breast (CC view)

(b) correctly classified dense breast (MLO view)

(c) incorrectly classified non-dense breast (CC view)

(d) incorrectly classified non-dense breast (MLO view)



(e) incorrectly classified dense breast (CC view)

(f) incorrectly classified dense breast (MLO view)

(g) correctly classified non-dense breast (CC view)

(h) correctly classified non-dense breast (MLO view)

Figure 4.5: Density classification results illustration. Example results obtained from one of the 40 repetitions.

Table 4.4: Selected screening classification results (not clustering the breasts according to density). Results correspond to mean values over 40 repetitions.

Classifier	Features		TPr	TNr	FNr	FPr
	Extraction technique	Type				
SVM RBF	squares	Gabor	0.9135	0.1034	0.0865	0.8966
		$GLCM_1$ $*$	**0.9267**	0.1382	**0.0733**	0.8618
		LBP	0.9158	0.1059	0.0842	0.8941
SVM sigmoid	breast	LBP	0.9189	0.1096	0.0811	0.8904
	rectangle	LBP	0.9113	0.1095	0.0887	0.8905
		Zernike	0.9013	0.1095	0.0987	0.8905
kNN	rectangle	InvMom $**$	0.9013	**0.1697**	0.0987	**0.8303**
	squares	Gabor	0.9194	0.1330	0.0806	0.8670

the idea of separating the images according to the tissue density for improving the detection of suspicious images.

If a supervised classifier is used the results are, as expected, not as good as those with the GT information. Nevertheless, when comparing the line marked with $*$ in Table 4.4 with the line marked with $*$ in Table 4.6, differences are significant for the True Positive (TP) and FN metrics.

Table 4.5: Selected screening classification results (using the GT information to cluster the breasts according to density). Results correspond to mean values over 40 repetitions.

Classifier	Features		TPr	TNr	FNr	FPr
	Extraction technique	Type				
SVM RBF	breast	$GLCM_1$	0.9147	0.1100	0.0853	0.8900
	rectangle	Gabor	0.9245	0.1305	0.0755	0.8695
		$GLCM_1$	0.9029	0.1206	0.0971	0.8794
	squares	Gabor	0.9213	0.1648	0.0787	0.8352
		$GLCM_1$	0.9004	0.1384	0.0996	0.8616
		$GLCM_2$	0.9126	0.1453	0.0874	0.8547
		Zernike $*$	**0.9404**	0.1010	**0.0596**	0.8990
SVM Sigmoid	rectangle	Gabor	0.9129	0.1473	0.0871	0.8527
		Zernike	0.9022	0.1222	0.0978	0.8778
kNN	breast	$GLCM_1$	0.9367	0.1158	0.0633	0.8842
		$GLCM_2$	0.9088	0.1284	0.0912	0.8716
		LBP	0.9132	0.1264	0.0868	0.8736
	rectangle	Gabor	0.9400	0.1527	0.0600	0.8473
		$GLCM_1$	0.9309	0.1037	0.0691	0.8963
		$GLCM_2$	0.9082	0.1119	0.0918	0.8881
	squares	Gabor $**$	0.9041	**0.2715**	0.0959	**0.7285**
		$GLCM_2$	0.9066	0.1690	0.0934	0.8310
		LBP	0.9231	0.1090	0.0769	0.8910

Table 4.6: Selected screening classification results (aggregating the breasts with a supervised classification method). Results correspond to mean values over 40 repetitions.

Classifier	Features		TPr	TNr	FNr	FPr
	Extraction technique	Type				
SVM RBF	squares	LBP	0.9054	0.1015	0.0946	0.8985
LDA	rectangle	$GLCM_2$	0.9025	0.1791	0.0975	0.8209
kNN	breast	LBP	0.9022	0.1545	0.0978	0.8455
		Zernike	0.9182	0.1035	0.0818	0.8965
	rectangle	Gabor $*$	**0.9518**	0.1099	**0.0482**	0.8901
	squares	$GLCM_2$ $**$	0.9196	**0.2143**	0.0804	**0.7857**

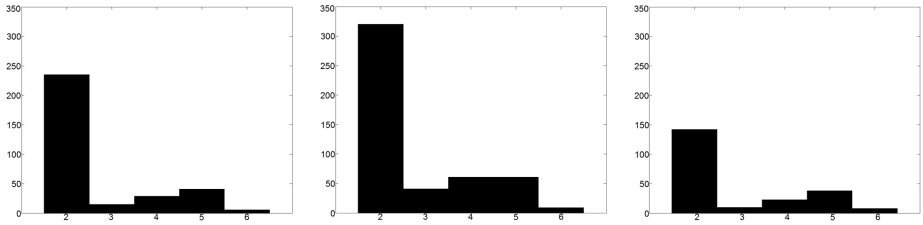
However, for the lines marked with $**$, the differences are not significant for any of the adopted metrics.

In conclusion, and since the method used to generate Table 4.5 needs user input and is thus not feasible in a practical setting, the method with supervised classification is the most appropriate. This method will be used in the remaining experiments.

Note that several other combinations could be tested, for example using different features for different tissue types, different classifiers for different tissue types, treating different views in different ways, etc. The hypothesis that separating the images according to the breast density improves classification results is, however, confirmed. This is valid as a proof of concept by using the GT information, and in a practical setting, by using a fully automatic method with previous breast density classification.

One further aspect that should be checked concerns the goal that no cancerous breast should be classified as non-suspicious. In order to check this, for every repetition, the real BI-RADS class of the images wrongly classified as non-suspicious was recorded. Histograms can be seen in Figure 4.6.

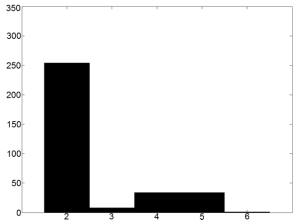
All the histograms are right skewed, meaning that most of the misclassified images have a low BI-RADS (and thus a low probability of being malign). Future directions to improve the



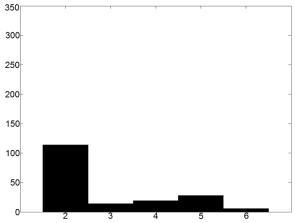
(a) SVM classifier with RBF kernel using LBP features extracted from squares inside the breast area

(b) LDA classifier with $GLCM_2$ features extracted from the biggest rectangle inside the breast region

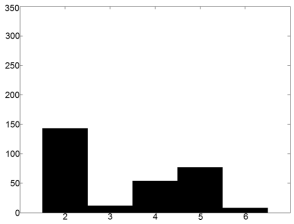
(c) kNN classifier with Zernike features extracted from the breast region



(d) kNN classifier with LBP features extracted from the breast region



(e) kNN classifier with Gabor features extracted from the biggest rectangle inside the breast region



(f) kNN classifier with $GLCM_2$ features extracted from squares inside the breast area

Figure 4.6: BI-RADS classes of the images wrongly classified as non-suspicious.

screening method include combining information from the two views of the same breast on the benign breasts identification block and/or combining several of the classification methodologies presented.

Recalling the application scenario described at the end of Section 4.1, where the system would replace one of the radiologists, it is interesting to compare the system performance with a human specialist. In Heywang-Koebrunner and Schreer (2014) it is stated that: "Realistically, mammography has a sensitivity of about 85%; that is, around 15% of carcinomas which are otherwise symptomatic at the time of the mammographic examination are not detected initially by mammography." As seen in Table 4.6, the automatic system can retrieve a sensitivity (or TPr) of more than 91% and FNr of about 8% or less. These values exceed those reported for human experts. The system could thus be used in practice as one of the two readers currently used in the screening pipeline.

4.5 Conclusions

This chapter describes an exploratory work for breast screening. In spite of the fact that the architecture used is not new, it is the first time that a fully automatic system has been built for screening purposes. A thorough evaluation by testing a large set of features in combination with

several classifiers was presented. Results have shown a sensitivity and a FNr better than those reported for human specialists. It could thus be used in a real clinical setting in combination with a human expert evaluation. A possible way to introduce an automatic screening system in the current practice is to replace one of the radiologists during the double-reading by the automatic decision system. In cases where a disagreement exists, the exam should be sent for further investigation.

Chapter 5

Detection of suspicious regions

Having screened out normal mammograms, the radiologist typically looks for suspicious regions in the mammogram. To assist the radiologist in this process, it would be helpful to have a system which can automatically identify possible regions of interest (ROIs) to be further analysed. There are mainly two types of findings that can be seen in mammogram images, calcifications and masses. Due to their differences, specialized detection systems are typically developed for each finding. In this way, this chapter is divided into two parts. In the first the detection of calcifications[1] is addressed, while the focus of the second part is on the detection of masses[2].

The main contributions of this chapter are: (1) the application of a Bayesian surprise technique to calcification detection, (2) the use of a new mass segmentation method needed for feature extraction in the false positives reduction step, and (3) evaluation of automatic detection techniques on the INbreast database.

5.1 Related work

The literature on suspicious regions detection in mammograms is extensive. Several review papers have already been published, both on the detection of calcifications (Cheng et al., 2003; Naqa and Yang, 2005; Rizzi et al., 2012), and the detection of masses (Oliver et al., 2010) or both (Karssemeijer and Hendriks, 1997; Sampat et al., 2005b; Rangayyan et al., 2007; Tang et al., 2009; Ganesan et al., 2013a; Jalalian et al., 2013). Here, only some selected recent works are reviewed.

5.1.1 Detection of calcifications

Naqa and Yang (2005) determine that calcification detection methods can be broadly separated into four categories, (1) basic image enhancement methods, (2) multi-scale decomposition methods, (3) stochastic modelling methods and (4) machine learning methods.

[1] Publications related to calcification detection include (Domingues and Cardoso, 2014; Tkaczuk et al., 2010).

[2] Publications related to mass detection and segmentation include (Kozegar et al., 2013; Domingues and Cardoso, 2013a; Cardoso et al., 2014).

However, most of the existing research uses hybrid approaches, where two or more of the above techniques are combined. Zhang et al. (2013), for instance, propose to first enhance the calcifications with a top-hat transform. Next, de-noising is performed using a wavelet decomposition. Finally, calcifications are detected based on their feature distributions. Experiments on a private database detected 92.9% true calcifications with an average of 0.08 false calcifications per image.

The work of Huang et al. (2013) uses three image filtering techniques in the pre-processing stage: top hat filtering to alleviate the uneven background problem; a wavelet transform to obtain the high frequency components; and Laws filters for further textural feature extraction. An iterative method is then executed, where morphological and edge detection operations are performed, to obtain a set of candidate regions. False Positives (FPs) are reduced using a Support Vector Machine (SVM) based classifier. Several texture and shape features are extracted and feature selection methods are compared. Results on a private database obtain a sensitivity of 92%, with an area under the curve of 0.99 and 0.65 FPs per image.

In his Doctoral Thesis, Torrent (2013) proposes a simultaneous detection and segmentation strategy for general objects that he then adjusts to the calcification detection case. The method begins by learning the variation in the morphology of the calcifications using local image features. Then, the set of features is used to train a pixel-based boosting classifier that, at each round, selects the most salient calcification feature. When a new mammogram is tested, only the salient features are computed and used to classify each pixel in the mammogram. Afterwards, the clusters are found by inspecting the local neighbourhood of each calcification. Experiments were made using the 322 digitized mammograms from the MIAS database and 280 digital mammograms from a private database. The area under the Receiver Operating Characteristic (ROC) curve was 0.903 for the MIAS database and 0.918 for the private database.

A deep learning technique is used in (Shin et al., 2014). After a local peak detection scheme, patches around the detected regions are manually classified as containing or not containing calcifications. A Discriminative Restricted Boltzmann Machine is then used to automatically learn calcification morphology and consequently classify new patches. Results using 9-fold cross validation on a private database of 33 mammograms reached an area under the ROC curve of 0.83.

Zhang et al. (2014a) propose a mathematical morphology and SVM method. First, the contrast in the original mammogram is improved by gamma correction and two structural elements are used to enhance the calcifications. Next, the potential regions are extracted using a dual-threshold technique. Finally, an SVM classifier is used to reduce the number of FPs. The performance of the proposed method is evaluated using the MIAS database. The experimental results achieved a TP rate of 94.85%, a FP rate of 7.82% and 0.53 FP calcifications per normal mammogram without calcifications.

In (Zhang, 2014) calcification detection is formulated as a sparse feature learning problem. A visual information-rich vocabulary of training samples is manually built from a set of samples, which include calcification parts and no-calcification parts. The sparse feature learning is acquired by the l_p-regularized least square approach with the interior-point method. Then the sparse feature

learning based calcification classification algorithm is designed using twin SVMs. To investigate its performance, the method is applied to the DDSM dataset. It achieves an average sensitivity of approximately 92.17% with respect to 7.83% FP rate and an area under the ROC curve of 0.9507.

Perhaps the most similar works to the one presented here are those that use novelty detection. The approach of Rose (2005) is motivated by the fact that signs indicative of breast cancer are not found in pathology-free mammograms. The approach requires a model of what normal mammograms look like. Rose's (2005) thesis presents two generative statistical models. The first treats mammographic appearance as a stationary texture. The second models the appearance of entire mammograms. Results in simulated calcifications achieved an area under the ROC curve of 0.92. The area under the ROC curve for images with simulated masses and calcifications was 0.75. When testing in images with real findings, the area dropped to 0.56 for images with calcifications only and to 0.53 for images with both masses and calcifications.

5.1.2 Detection of masses

Oliver et al. (2010) classify mass detection methods according to the flowchart in Figure 5.1. In

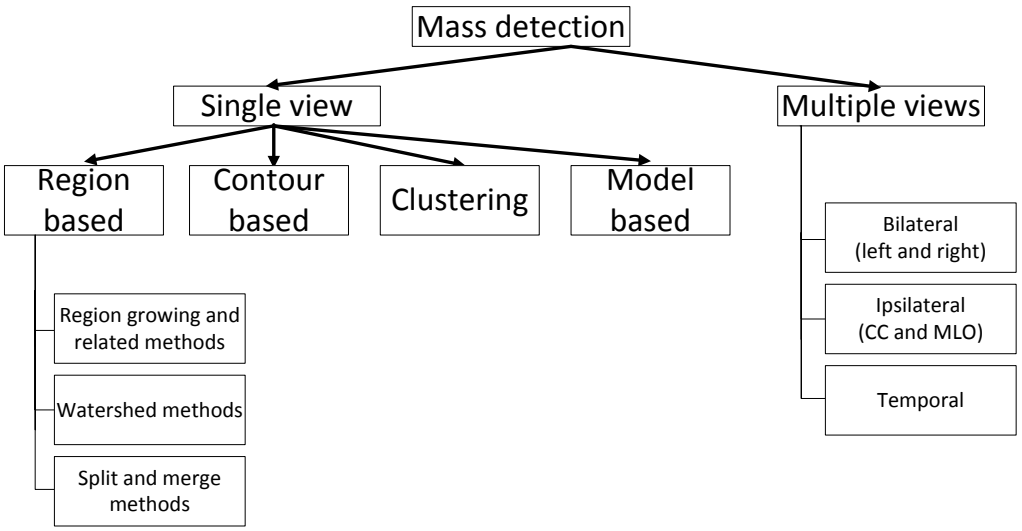


Figure 5.1: Classification of mass detection methods, based on (Oliver et al., 2010).

order to study how the performance varies for the different strategies that use one single view, they have quantitatively compared seven different approaches based on: (1) detection of concentric layers, (2) Laplacian edge detector, (3) thresholding, (4) Iris filter, (5) difference of Gaussians, (6) pattern matching, and (7) a classifier approach. The algorithms have been evaluated using two different databases: MIAS and a Full Field Digital Mammography (FFDM) private database. The best overall results were attained with the Iris filter approach, with an area under the ROC curve of 0.787 for MIAS and 0.757 for the FFDM private database.

A more recent example of a single view method is the one in Agrawal et al. (2014) and Agrawal (2014). The framework uses saliency based segmentation which does not require removal of pectoral muscles, if present in the mammogram. Graph Based Visual Saliency computes the saliency of a region with respect to its local neighbourhood using the directional contrast. The steps involved are: (1) computation of feature maps from contrast values along four different orientations of 2D Gabor filters (0^o, 45^o, 90^o, and 135^o); (2) computation of activation maps; (3) normalization of activation maps; and (4) combination of the normalized activation maps using the sum rule. Segmentation is obtained by thresholding the obtained saliency map. From the segmented regions, different features are extracted followed by SVM classification for mass detection. The experiments are performed using the MIAS database. 49 out of a total of 58 masses are detected with $2-3$ FP/image.

Ericeira et al. (2013) propose a multi-view scheme where bilateral information is used. Asymmetric regions between pairs of mammograms of the left and right breast are detected by means of structural variations between corresponding regions, defined by a spatial descriptor called a *cross-variogram* function. After determining the asymmetric regions of a pair of images, the variogram function is applied to each asymmetric region separately for classification as either mass or non-mass. Results on the Digital Database for Screening Mammography (DDSM) database were: 90.26% of accuracy, 100% sensitivity and 85.37% specificity.

Ipsilateral information is used in the work of Pereira et al. (2014). A pre-processing method based on wavelet transform and Wiener filtering is first applied for image de-noising and enhancement. A Genetic algorithm is then employed for segmentation of suspicious regions. A post processing step where the area of the structures marked on the CC view is manually compared to that on the MLO view is finally performed. A FP rate of 1.35 FP/image is obtained for a sensitivity of 95% using DDSM database. The post-processing procedure is, however, not completely automatic.

In comparison with the other multiple view methods, very little research has been done using temporal information. Unfortunately, similar to DDSM and MIAS, the INbreast database also does not allow for these types of studies.

5.1.3 False positive reduction

The main issue with many mass and calcification detection methods is that, in order to minimize the number of missed findings, a high number of FPs is produced. Thus, some work focuses on the reduction of the FPs. This section summarizes some of the recent contributions towards reducing the number of FPs.

Nguyen et al. (2013) use Block Variation of Local Correlation Coefficients texture features to characterize mass patches. Then, SVMs are used for classification. Evaluation on 2 700 ROIs from the mini-MIAS database gives an area under the ROC curve of 0.93 and FPs reduce 82% in relation to the baseline detection.

In (Hussain et al., 2014) the FP reduction problem is addressed using textural properties. A Gabor filter bank is used to extract the most representative and discriminative textural properties of

masses present at different orientations and scales. An SVM with a Gaussian kernel is employed for classification. The method is evaluated over 1 024 (512 masses and 512 normal) ROIs extracted from the DDSM database. The best result (area under the ROC curve $= 0.96 \pm 0.02$) is obtained when a Gabor filter bank with 5 orientations and 3 scales and ROIs with size 512×512 is used.

Lavanya et al. (2014) propose a feature-fusion-based system for automatically classifying detected masses as true masses or FPs. In their system, unilateral and bilateral information is fused using Canonical Correlation Analysis (CCA). The system is validated using the MIAS database. The overall classification performance is higher by a range of $8\% - 16\%$, $12\% - 16\%$ and $14\% - 28\%$ in terms of accuracy, area under the ROC curve and equal error rate, respectively. The reduction in FP rate is at least 39%, 35% and 33% at TP rates of 60%, 65% and 70%, respectively.

Vallez et al. (2013) observe that the performance of a CAD system can be improved if breast density information is considered. They conclude that it is necessary to adjust the input parameters of a lesion detection algorithm to control its sensitivity depending on the tissue type in order to reduce FP detections. Thus, prior to the detection algorithm, tissue classification is applied. Both a private and the mini-MIAS databases were used in the experiments and evaluation was made both for mass and calcifications detection. On average the TPs increased 13% and the FPs decreased 14% with tissue type classification. The overall sensitivity obtained before tissue type classification was 77.5% and after 90.5%. The overall specificity obtained before tissue type classification was 77.9% and after 91.6%.

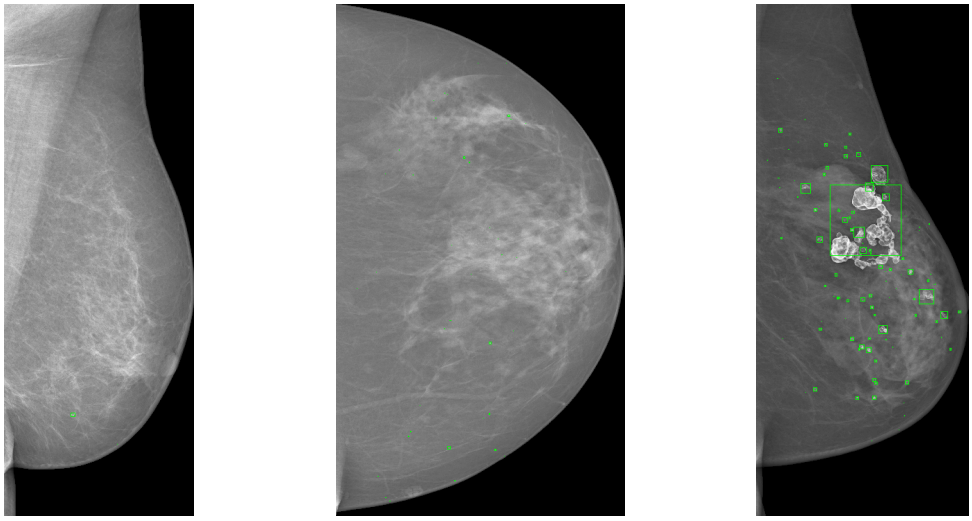
As can be seen, FP reduction techniques are typically based on machine learning approaches where a set of features is first extracted then a two-class classification methodology is used. Most of the works present results using databases of "synthetically" generated patches. In spite of the fact that they are real examples, they are not produced by mass detection systems, but instead generated using Ground Truth (GT) information. In this way, it is not possible to calculate the impact of the FP reduction techniques in a real scenario. Some exceptions include (Nguyen et al., 2013; Vallez et al., 2013) where patches generated from a detection system were used.

5.1.4 Evaluation

When judging experimental results and comparing the relative performance of different methods, it is important to realize that the definitions used for TPs and FPs often differ. In this way, even two methods evaluated on the same dataset may not be directly comparable. The most common ways to check whether a suspicious area marked by a system corresponds with an annotated lesion are: (1) a region is considered as detected if the centre of gravity of the automatic annotation is inside the manual annotation; (2) if the pixel with the highest measure of suspiciousness falls inside the manual annotation, (3) using an overlap criterion (Oliver et al., 2010). Karssemeijer and Hendriks (1997) found that different criteria gave similar results when the false alarm rate is low, but results may differ significantly when there are more than three FPs per image.

5.2 Detection of calcifications

The INbreast database contains calcifications with areas ranging from 1 pixel^2 to more than 20 000 pixel^2. Figure 5.2 shows the diversity of calcifications that may be found in mammogram images.



(a) image with one calcification only

(b) image with several calcifications

(c) image with large calcifications

Figure 5.2: Examples of calcifications.

The method here proposed to detected calcifications is inspired by Bayesian surprise (Section 5.2.1). As a high number of FPs is generated by the detection technique, an FP reduction methodology is subsequently used (Section 5.2.2).

5.2.1 Bayesian surprise detection

The surprise caused by an observation is defined in a Bayesian sense as the change it brings to an observer's prior beliefs with respect to the phenomenon under consideration (Gkioulekas et al., 2010). It can be mathematically defined as follows (Itti and Baldi, 2005): given a prior distribution $P(M)$ over a (discrete) space of models $\mathscr{M}$ describing a phenomenon, and the posterior distribution $P(M|D)$ after new data D is obtained for this phenomenon through an observation, the surprise incurred by D relative to the space $\mathscr{M}$ is given by the Kullback-Leibler divergence (K) between the prior and the posterior distribution,

$$S(D,\mathscr{M}) = K(P(M)||P(M|D)) = \sum_{M\in\mathscr{M}} P(M) log \frac{P(M)}{P(M|D)}. \tag{5.1}$$

Bayesian surprise can be used for images to explain the saliency of regions that, compared to the rest of the image, exhibit irregular characteristics (Gkioulekas et al., 2010). This is particularly

interesting for calcification detection since they usually correspond to bright spots in the mammogram image. The visual context of the region implies a prior distribution $P(M)$ over the model space before the region is observed. After the region is observed, a posterior distribution $P(M|D)$ is formed, where D is the data acquired from the observation of the region. The surprise incurred by D relative to the space $\mathscr{M}$ is then given by (5.1) (Gkioulekas et al., 2010). This interpretation is demonstrated in Fig. 5.3, where the area surrounding a square region serves as its context.

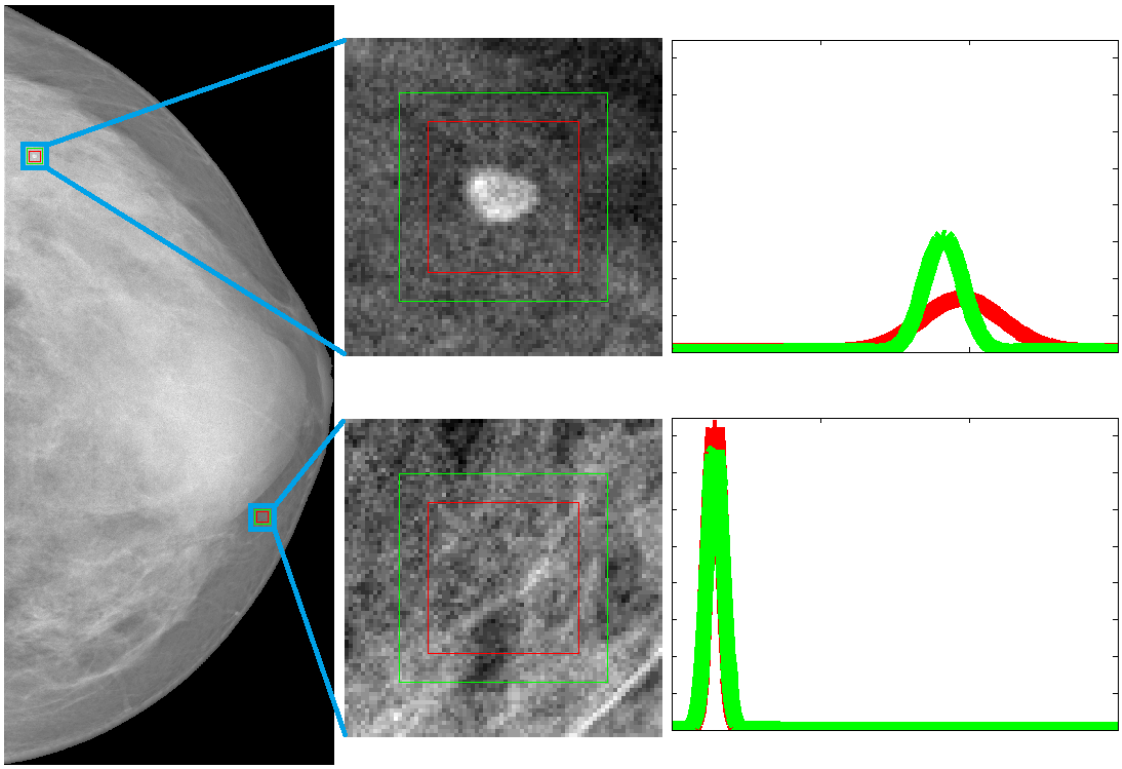


Figure 5.3: Spatial interpretation of surprise for two regions (red squares) and their respective context (frames between red and green squares). Intensity distributions are also shown (for better visualization purposes only, distributions were approximated by a Gaussian distribution in this illustration).

The methodology for applying this technique to calcification detection is:

- for each patch of the image
  - compute the surprise induced by the patch in relation to its neighbourhood
    - if the surprise value exceeds a threshold
      - consider the region as a calcification

The technique is repeated in a set of 10 scales, by extracting patches of sizes ranging from 12×12 to 372×372 pixels. The context region was defined as the square frame having the same area as the patch. In order to make the process less computationally expensive, instead of using a dense grid, patches with no overlap were extracted. As there are, generally, more small calcifications

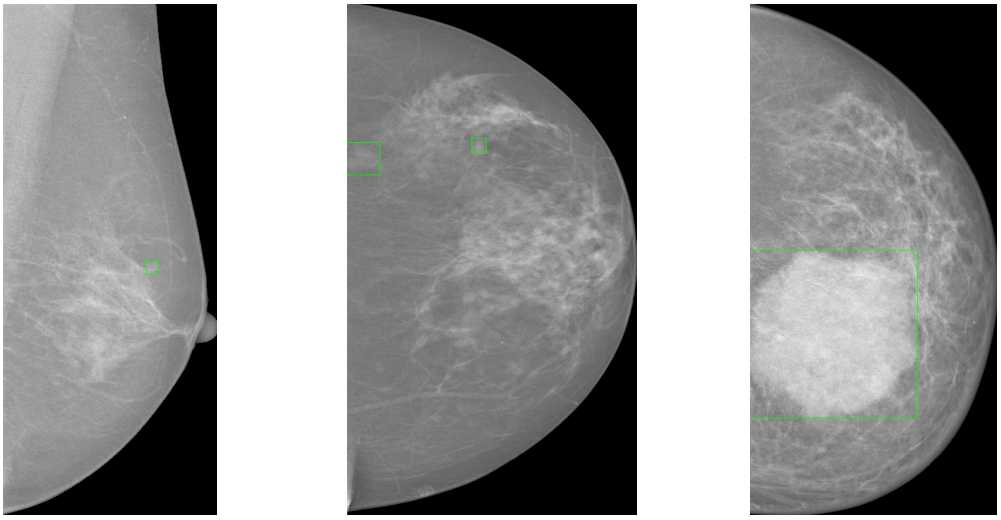
than larger ones, surprise threshold values were set according to a linear relationship: $Surp_{th} = m \times Patch_{width} + b$. Parameters were set experimentally to $m = 25$ and $b = 200$. If, for a particular image, the number of detected calcifications exceeds a predefined value ($NrDetectionsMax$), b is iteratively increased by 100 until the point when fewer than $NrDetectionsMax$ calcifications are detected.

5.2.2 False positive reduction

The above described detection technique tends to return a high number of FPs. To reduce the number of FP detections, a second step is performed, where some simple features are initially extracted and used in a classification algorithm. The feature vector of size 8 consists of the following characteristics: intensity value of the image at the detected point; standard deviation, minimum value, 25^{th} percentile, median value, mean value, 75^{th} percentile and the maximum value of the image intensities in the $Patch_{width} \times Patch_{width}$ patch around the detected value. An SVM with the Radial Basis Function (RBF) kernel was trained in a randomly selected subset containing 75% of the images in the dataset. This process is repeated 40 times in order to achieve more stable results.

5.3 Detection of masses

The INbreast database contains masses with areas ranging from 3 096 pixel^2 to 752 766 pixel^2. Figure 5.4 exemplifies the diversity of masses that may be found in mammogram images.



(a) image with small size mass **(b)** image with two masses **(c)** image with large mass

Figure 5.4: Examples of masses.

The steps involved in the mass detection algorithm are shown in the pipeline of Figure 5.5 and briefly described in the following paragraphs.

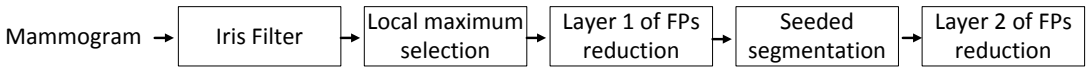
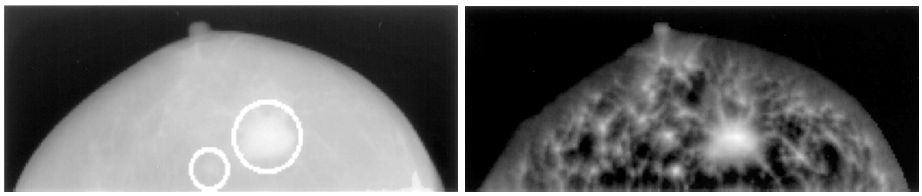


Figure 5.5: Mass detection pipeline.

As referred in Section 5.1.2, the Iris filter has proved to be one of the most successful approaches for mass detection. In brief, the Iris filter enhances rounded opacities by means of the analysis of a gradient-orientation map (Kobatake et al., 1998; Kobatake, 2006), as shown in Figure 5.6. The procedure often succeeds in cases of very low contrast because it depends only on the orientation of the gradient vectors in the image and not on their amplitude (Mencattini et al., 2009).



(a) original mammogram **(b)** output of the Iris filter

Figure 5.6: Iris filter example (Kobatake, 2006).

The result of this filter is that rounded opacities (such as masses) are highlighted, while linear structures are mostly removed (Oliver et al., 2010). The filter returns a pseudo-probability image in the range $[-1,1]$, where higher values are obtained near the centre of round convex regions (Oliver et al., 2010). Hence, thresholding these images at different levels leads to different levels of sensitivity and number of FPs.

In order to reduce the number of FP detections, some features are extracted from patches around each detected pixel. The **feature** vector of size 10 consists of the following characteristics: from the original images – intensity value at the detected point, standard deviation (*std*), 25^{th} percentile, median value, mean value, 75^{th} percentile, and maximum intensity value in the 10×10 patch around the detected value were extracted; from the Iris filtered images – the 25^{th} percentile, median value, and maximum value in the 10×10 patch around the detected value were extracted. SVMs with the RBF kernel were trained on randomly selected examples from 75% of the dataset and the process was repeated 40 times in order to achieve more stable results. The patch size was empirically defined, other values (including different sizes for the original images and the Iris filtered images) should be tested in the future.

The detected points that are not removed by the FP reduction technique are used as seeds for a segmentation technique. The segmentation method used is the closed shortest path (this method will be referred to as "CaPTOR") described in (Oliveira, 2013, Ch.5). In the adopted segmentation technique, the closed contour is computed in the original coordinate space, contrary

to traditional shortest patch segmentation methods where the images are first transformed using polar coordinates.

Note that this technique could have been used for detection instead of the Iris filter. CaPTOR is, however, computationally more expensive but provides more accurate segmentations. It was thus decided to use the Iris filter in the detection phase, followed by CaPTOR for segmentation. More details on the CaPTOR technique and its evaluation when segmenting masses are given in Sections 6.2.2 and 6.3.3.

After segmentation, a second set of **features** is extracted with the aim of further reducing the number of FPs. This vector consists of 9 shape features (area of the segmented region, area of the bounding box of the region, area of the region's convex hull, eccentricity of the ellipse that has the same second central moments as the region, length of the major axis of the ellipse that has the same normalized second-moments as the region, length of the minor axis of the ellipse that has the same normalized second-moments as the region, diameter of a circle with the same area as the region, orientation, perimeter), and a feature that uses both shape and intensity information (Euclidean distance between the centroid and the weighted centroid). A new SVM classifier was built as described for the first level of FP reduction.

It is worthwhile to note that the Iris filter and the first layer of FP reduction are made on resized images of size 512×256 (in order to speed up the process), while the remaining steps are performed using the original image size.

5.4 Evaluation methodology

As seen in Section 5.1.4, most of the evaluation methodologies individually look at each lesion in the GT and find a corresponding lesion on the detection results. There is, however, a significant drawback with this approach, since it does not force a one-to-one relationship and several detected regions may be wrongly associated to the same GT lesion. Here, the approach in (Moreira et al., 2012) is followed. The main steps are as follows:

- Dissimilarity Matrix computation: calculation of all the costs between the GT findings and the detected regions;
- Optimal assignment calculation: assignment of exactly one GT region to one and only one detected region in such a way that the total cost of the assignment is minimized (in the current implementation, the Hungarian Algorithm (Kuhn, 1955) is used).

For the calcification detection case, as they are small structures, they can be approximated by the corresponding centroid. In this way, the most natural dissimilarity metric to use is the Euclidean distance. For masses, as it is not reasonable to do a singular point approximation, the adopted metric is $(1 - AOM)$ where AOM is the amount of overlap between the GT region and the detected region (as defined in Section 2.8). A saturation threshold is then applied to the dissimilarity matrix in order to minimize incorrect assignments (see (Monteiro, 2011) for an illustration).

After the optimal assignment calculation, the number of FPs is automatically determined by observing the detected regions with no corresponding GT region. When dealing with clusters, a final stage is added where the FP detections are checked to see if they fall inside the GT cluster region. If they do, they are no longer considered FPs. The missed detection cases are assessed by setting an upper threshold in the allowed dissimilarity value.

5.5 Results

In the experiments made in this chapter it is assumed that: orientation homogenization has been performed, the breast region has been detected, and images with Breast Imaging Reporting And Data System (BI-RADS) class 1 have been screened out. For INbreast, this give a total of 343 images with the tags as shown in the Venn diagram of Figure 5.7.

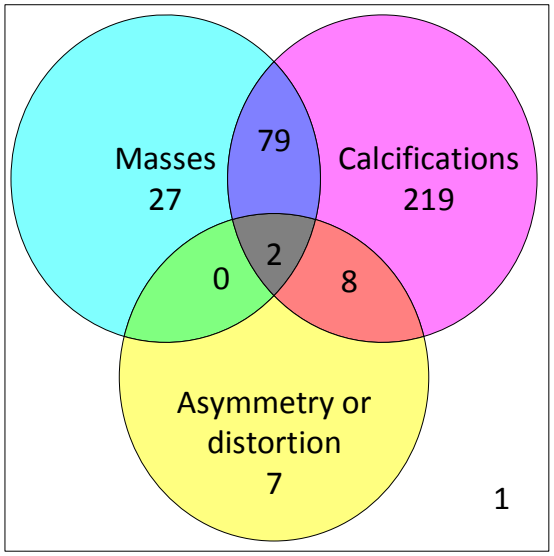


Figure 5.7: Distribution of findings in the INbreast database.

In the case of calcifications detection, it is also assumed that the pectoral muscle has been extracted from the MedioLateral Oblique (MLO) images (using the GT information). In the case of mass detection, the muscle was included in the analysis because it was observed that some of the INbreast mammograms contained masses in this region (Figure 5.8).

Sections 5.5.1 and 5.5.2 summarize the results on the detection of calcifications and masses, respectively.

5.5.1 Calcification detection

For comparison with the proposed Bayesian surprise detection method, other techniques were tested:

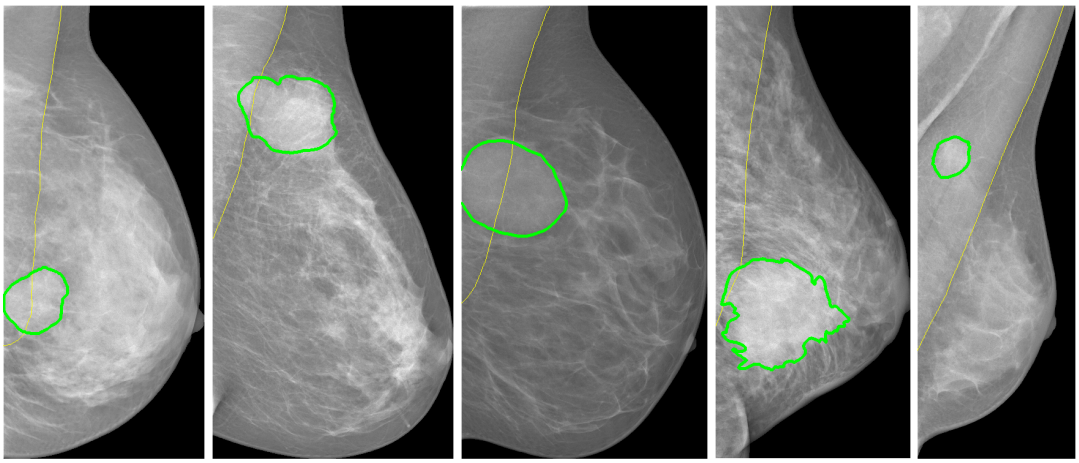


Figure 5.8: Examples of masses seen in the pectoral muscle region. The pectoral muscle is delineated in yellow while the mass contour is shown in green.

Fixed threshold: As calcifications are usually brighter than the remaining breast tissue, a first naive approach is to determine a fixed value (threshold) and classify all image pixels with intensity above that value as suspicious. The problem resides in the determination of a unique robust threshold value that works for all the images.

Outlier detection: Instead of using the same threshold for all images, it can be adjusted individually for each image. In this technique, all points with intensity higher than $q_3 + 1.5 \times (q_3 - q_1)$ where q_1 and q_3 are the 25^{th} and 75^{th} percentiles, are considered as outliers and thus as calcifications. The value 1.5 corresponds to approximately 99.3% coverage if the data (i.e. pixel intensities) are normally distributed (Velleman and Hoaglin, 1981).

Mathematical morphology: For comparison with the state of the art, the technique of Zhang et al. (2014a) based on mathematical morphology was reimplemented. The FP reduction part was not performed since, as will be shown in the experimental section, this method has a low sensitivity. Reducing the FPs would decrease the sensitivity even more.

From Table 5.1 it can be seen that the method with the worst sensitivity is the Fixed threshold, followed by Mathematical morphology, Outlier detection and the highest sensitivity is achieved with the Bayesian surprise. Concerning the FPs, the method with the worst behaviour is Bayesian surprise, followed by Mathematical morphology, Fixed threshold and Outlier detection.

Table 5.1: Calcification detection results

Method	Sensitivity (%)	FP
Fixed threshold	28.2	67
Outlier detection	44.9	52
Mathematical morphology	29.3	78
Bayesian surprise	58.2	96

A comparative plot of the performances in several operating points can be seen in Fig. 5.9. The different behaviours were accessed by: (1) in the Fixed threshold method, the threshold was varied

between 1.0 and 0.8; (2) for the Outlier detection technique the computed individual threshold was varied by adding a constant between -0.50 and 0.50; (3) the Mathematical morphology method has two parameters, α and β (see Zhang et al. (2014a)) which were varied in the interval between 0.0 and 1.0; (4) finally, in the Bayesian surprise, the number of maximum detections was varied between 0 and 400.

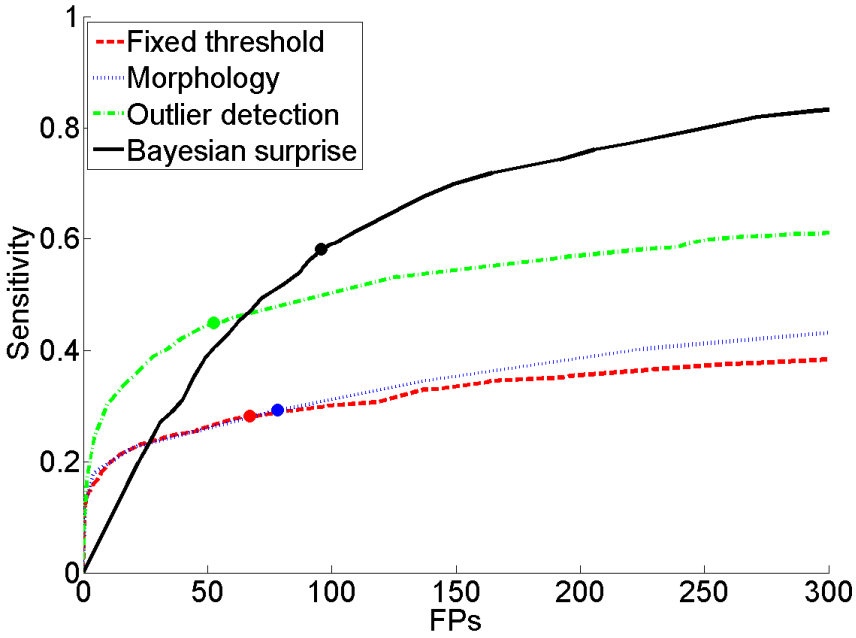


Figure 5.9: Comparative plot of the performances of the detection methods. Circles correspond to the operating points in Table 5.1.

The curves show that Bayesian surprise has a poor performance for small sensitivity values but it rapidly performs better than all the other tested methods as sensitivity increases. Illustrative examples are presented in Fig. 5.10.

To test the FP reduction, 40 different splits into train/test were made. When no FP reduction is performed, the method achieves a Sensitivity of $66.6 \pm 3.5\%$ with an average number of 70 ± 2 FP detections per image. After the reduction of FPs, the sensitivity of the Bayesian surprise method decreases to $51.9 \pm 7.9\%$ and FP to 37 ± 11. Some examples are shown in Fig. 5.11.

Although a direct comparison with the results in Table 2.4 is not possible, due to differences in the databases, GT specification methodology, evaluation technique, etc., the results for the Bayesian Surprise method compare well in terms of sensitivity with the commercially available approaches.

5.5.2 Mass detection

The main parameter of the Iris filter is the radius of the circular neighbourhood for each point. Three different radii were tested, 10, 15 and 20, leading to three filtered images for each mammogram. Next, the top N local maxima of each filtered image were calculated. Finally, the detected

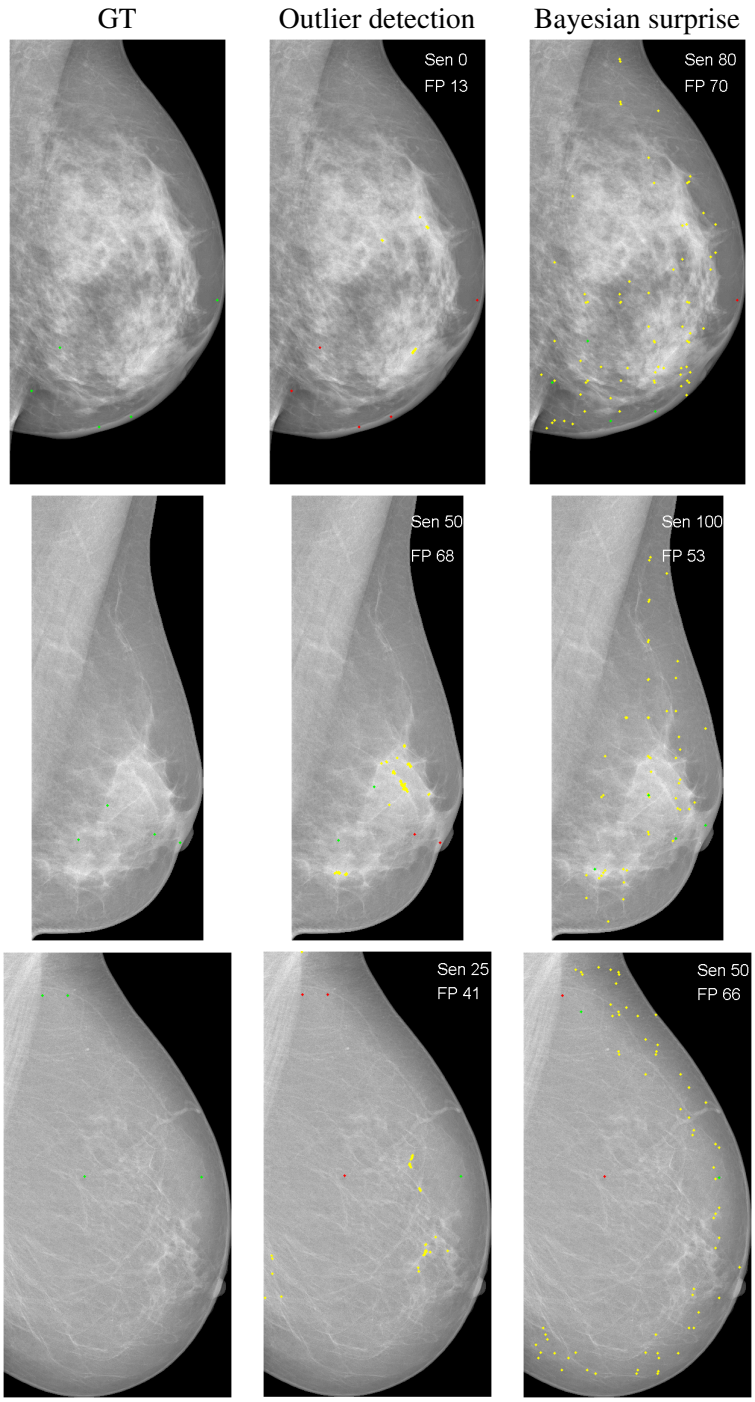


Figure 5.10: Examples of calcification detection. Green dots correspond to TPs, yellow to FPs and red to FNs. Sensitivity and number of FPs are also shown.

points for each radius were merged and the initial suspicious regions were retrieved. Examples of detections using different N values are given in Figure 5.12.

The impact of the first and second layers of FP reduction can be seen in Figure 5.13. As the combination $N = (0,0,40)$ subjectively seemed to be the most promising one, by detecting an important amount of masses at a reasonable FP rate, only results for that case are presented.

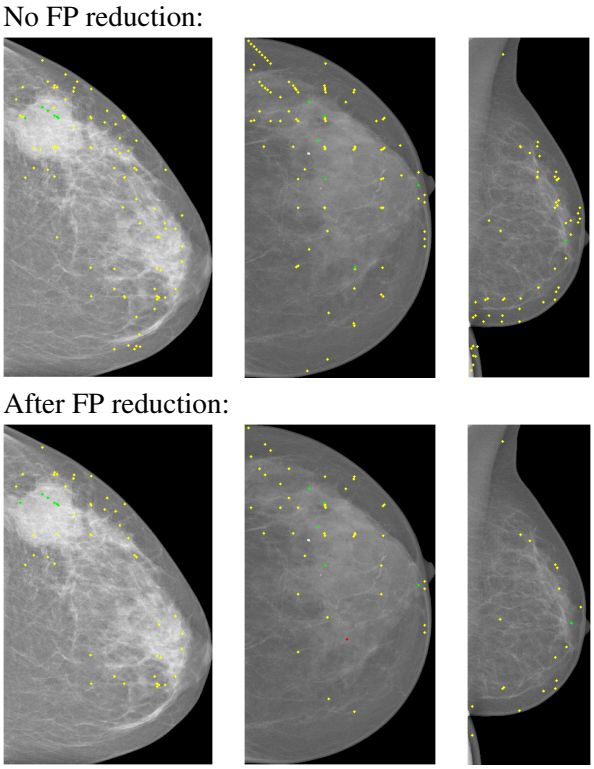


Figure 5.11: Examples of calcification detection before and after FP reduction. Green dots correspond to TPs, yellow to FPs and red to FNs.

The overall performance for all 343 images of the mass detection scheme is: Sensitivity $= 38\%$ (2.5) with FP $= 5$ (0.2). It is interesting, however, to see the impact of some factors in the detection. The first factor to be studied is breast density. Results are summarized in Table 5.2. Conducting pairwise statistic tests across each combination of American College of Radiology (ACR) densities, differences are shown to be statistically significant. This decrease of mass detection sensitivity with the increase of breast density has also been noted by other authors, e.g. (Garcia-Manso et al., 2013).

Table 5.2: Mass detection results according to breast density. Results are presented in the format: mean (*std*).

ACR density	Total number of images	Sensitivity	FP
I	114	52% (4.5)	3 (0.2)
II	114	30% (3.8)	3 (0.2)
III	89	36% (4.3)	6 (0.4)
IV	26	7% (6.2)	9 (0.5)

The split of the results according to BI-RADS is shown Table 5.3. Sensitivity for BI-RADS 4 is shown to be significantly better than the sensitivities of all the other BI-RADS classes. This is a surprising result since it is well known that BI-RADS category 4 lesions are associated with a highly variable rate of breast cancer (risk of malignancy between 2% and 95%) (Flowers et al., 2013). Since other systems have also shown behave differently according to the BI-RADS categories (see, for instance the work (Ellis et al., 2007) where both ImageChecker and SecondLook

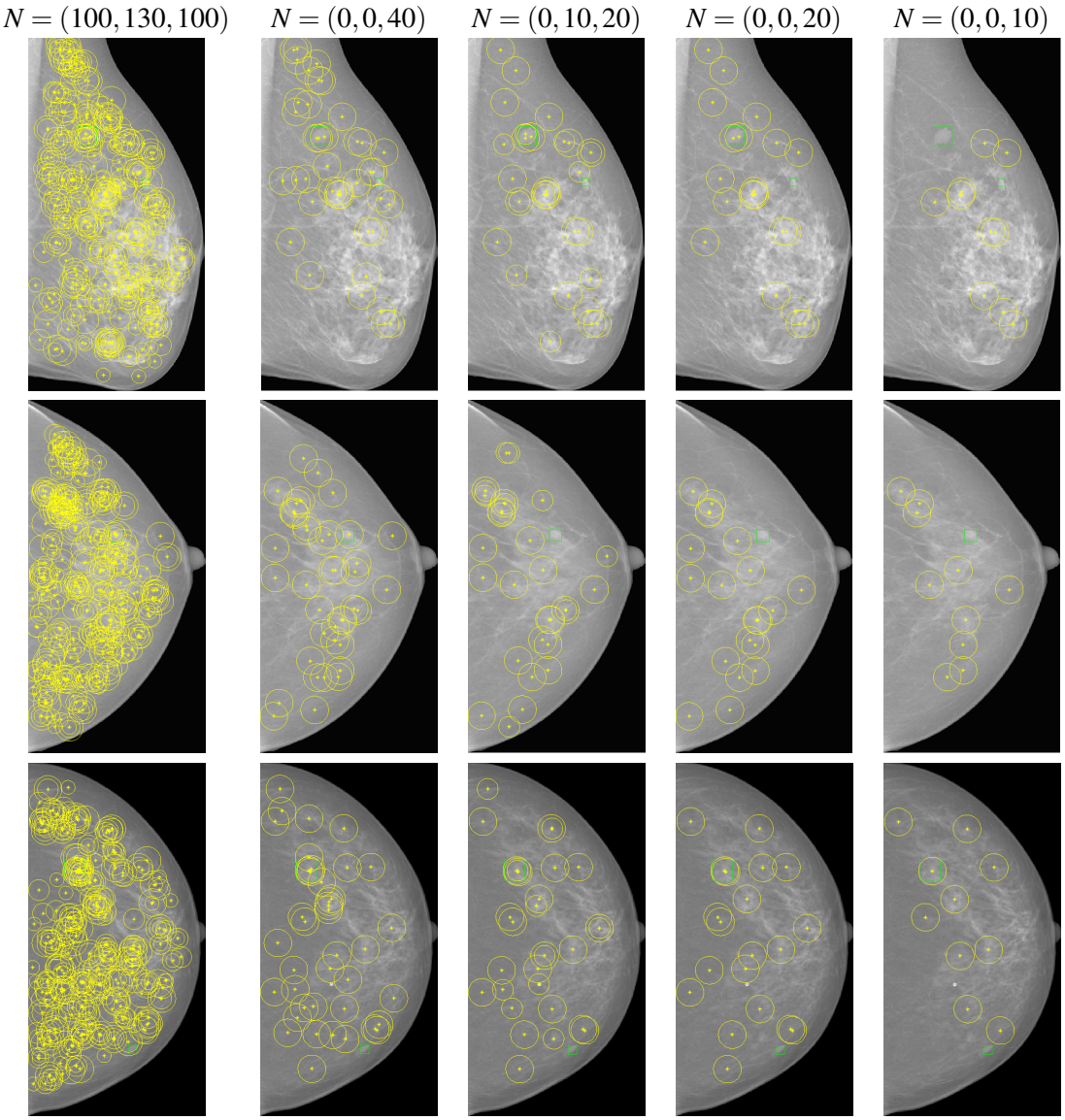


Figure 5.12: Examples of mass detection. Detected regions are shown in yellow while a bounding box around GT annotations are shown in green. N vector corresponds to (N_{10}, N_{15}, N_{20}), where N_i is the N parameter used in the image filtered by an Iris filter of radius i.

presented better results for category 5 than category 4), a possible future work direction is to develop different systems for different BI-RADS categories. A drawback of this approach is that the BI-RADS assessment must be known before the mass detection stage.

Table 5.3: Mass detection results according to BI-RADS. Results are presented in the format: mean (*std*).

BI-RADS	Total number of images	Sensitivity	FP
2	220	31% (2.7)	5 (0.2)
3	23	31% (7.2)	7 (0.3)
4	43	59% (6.6)	6 (0.3)
5	49	34% (3.6)	5 (0.3)
6	8	35% (11.1)	5 (0.5)

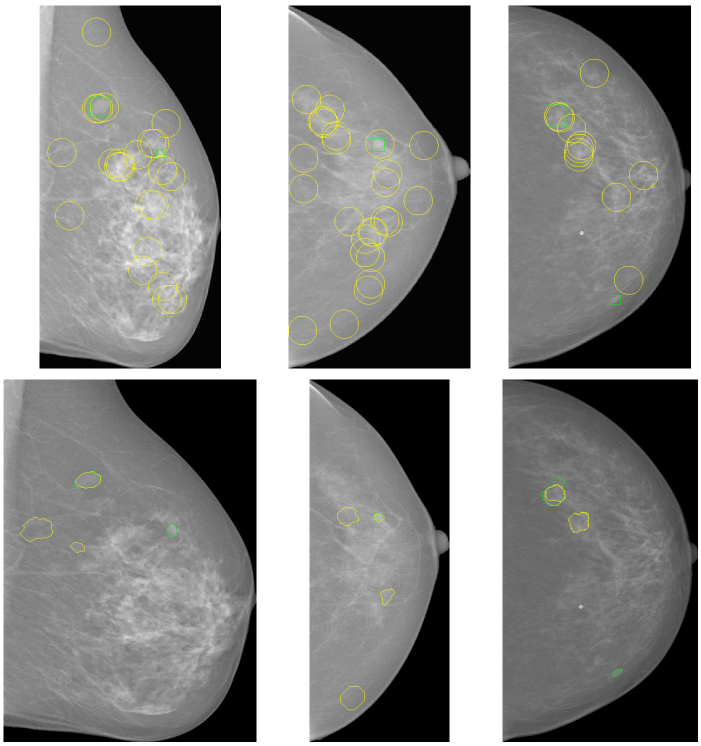


Figure 5.13: Examples of Mass detection after FP reduction for one of the 40 repetitions. Top row illustrates the results after the first layer of FP reduction and bottom row corresponds to the second layer of FP reduction. Differences in image sizes are due to the fact that the first part of the process is made in re-sized images while the second part is made using the original image.

The effect of mass size on the performance of detection algorithms has also been studied. While in (Bargallo et al., 2013; Murakami et al., 2013) no influence of lesion size on the sensitivity was found, in (Malich et al., 2003) the highest detection rates were observed for 10-30 mm tumor masses. For the presented algorithm (Table 5.4), detection is significantly better for the 10 to 20 mm range and worst for masses larger than 20 mm. These results are likely to be a direct consequence of the radius parameter chosen for the Iris filter. While some preliminary experiments of combining results using different radii were attempted (recall Figure 5.12), this needs further exploration in future work.

Table 5.4: Mass detection results according to mass size. Results are presented in the format: mean (*std*).

Mass size (mm)	Total number of masses	TP
≤ 10	32	32% (5)
10 to 20	43	65% (4)
> 20	40	13% (3)

Another important factor that may affect the performance is mass spicularity (Table 5.5). Conducting pairwise statistic tests between non-spiculated and spiculated results, differences on sensitivity and FPs are shown to be significant. The algorithm is thus better at detecting non-spiculated masses than at detecting spiculated masses. While both human experts, e.g. (Rawashdeh et al.,

2013), and some of the available techniques also show the same behaviour, e.g. (Juarez-Landin et al., 2008), other methods have been proposed to deal with the detection of spiculated masses, e.g. (Sampat et al., 2008). An algorithm that combines one method for non-spiculated masses and a method for spiculated masses is a promising area for future research.

Table 5.5: Mass detection results according to mass spicularity. Results are presented in the format: mean (*std*).

Mass type	Total number of images	Sensitivity	FP
non-spiculated	91	40% (2.7)	6 (0.2)
spiculated	16	34% (5.5)	4 (0.3)
overall	107	38% (2.6)	6 (0.2)

In more recent work, done in collaboration with an international team (Kozegar et al., 2013) better results were achieved. There, masses are detected in two main steps: (1) extraction of suspicious regions using an adaptive threshold and (2) false positive reduction. Assessment on the 107 images with masses from INbreast achieved a sensitivity of 87% (with 15 missed masses) and a FPr of 3.67 per image. It is important to note that, although the same database has been used, the evaluation methodology differs from the one used in the present work. The results are thus not directly comparable. Moreover, the results in (Kozegar et al., 2013) point to the fact that it is possible to develop better algorithms for mass detection or to improve the algorithm here proposed.

Finally, it should be noted that some of the FP detections, while not being masses, might correspond to other types of interesting findings such as calcifications (Figure 5.14).

5.6 Conclusions

In this chapter methods for the detection of the two most common mammogram findings were attempted. For the calcifications, a methodology not yet used for this application was suggested. For mass detection a state of the art technique was used as main processing block, followed by a machine learning technique for false positive reduction. Features given as input to the classifier were extracted from an automatic contour generated with a new segmentation technique that calculates the shortest path in the original coordinate space. Another contribution of this chapter was to provide baseline detection results over the INbreast database. To the best of our knowledge, no other work has yet been made on the detection of calcifications in this database. For masses, only one other work has been presented. Future proposals can thus be validated by comparison with the results presented here.

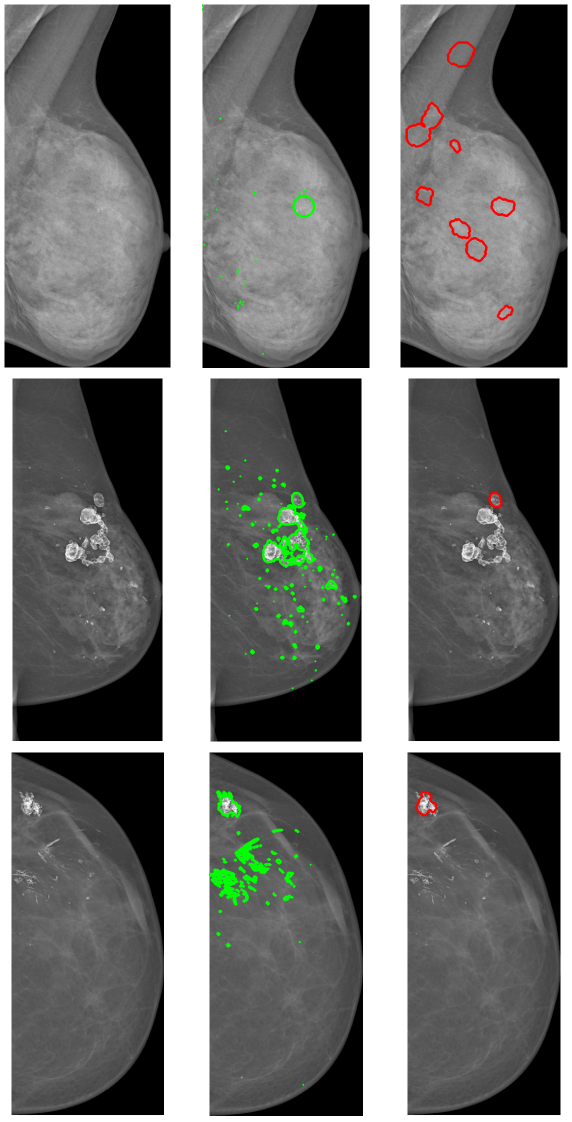


Figure 5.14: Examples of mass detection results that detect calcifications. Left: original mammogram; centre: GT contours; right: mass detection result.

Chapter 6

Characterization of calcifications and masses

The Breast Imaging Reporting And Data System (BI-RADS) standard recommends the description of calcifications according to their distribution and morphology, while masses are characterized through their margins, shape and density characteristics (Figure 6.1).

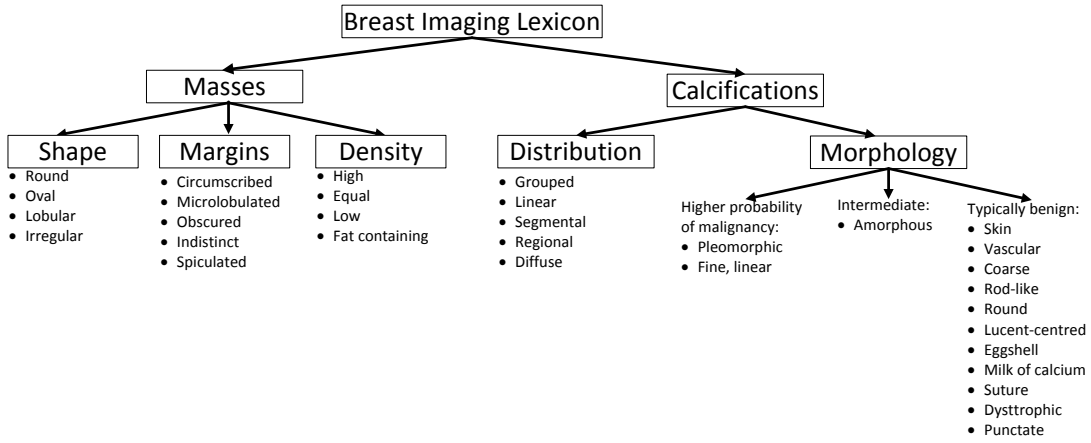


Figure 6.1: Hierarchical structure of the BI-RADS lexicon.

Even with its standardization and systematic approach, BI-RADS lacks guidance on how to determine the appropriate BI-RADS category assignments (Table 2.3) using the combination of BI-RADS descriptors in Figure 6.1 (Youk et al., 2012). Thus, most of the published works present methods for binary classification (e.g. benign/malign) only. It is, however, in the interest of this work to classify the images according to the BI-RADS categories in Table 2.3. This classification, besides being more natural to the specialists, also provides a degree of probability of malignancy. To the best of our knowledge, besides the framework presented in Chapter 7, there is no other published work that attempts to do this ordinal multi-class classification of the full mammogram.

There are, however, some works that classify specific findings according to some of the characteristics in the BI-RADS standard. In this chapter[1], a set of features is first compiled from the literature focusing on the potential features that may represent the findings as described in the standard. The remainder of the chapter consists of an exploratory study of these features. The features shown to be meaningful for the BI-RADS classification will then be used as input to the proposed classification method described in Chapter 7.

The main contributions of this chapter are the extraction of features from the INbreast database findings and the study of their relation to each BI-RADS class. The evaluation of a new segmentation algorithm for the mass delineation is also presented, along with the impact of the automatic contour on feature quality. Feature selection is performed with the Pearson correlation (*corr*) and with two other metrics not used before in the mammography context namely, distance correlation (*dcorr*) and Maximal information coefficient (MIC).

6.1 Related work

A literature review on the use of automatic features for the characterization of both calcification and masses is now presented. Special focus is given to works that try to extract features that represent the BI-RADS standard. However, as there is a very limited number of works on this topic, some other papers that attempt classification into benign/malign are also included. The first part of the review covers features used to characterize calcifications (Section 6.1.1) while the second part addresses features to characterize masses (Section 6.1.2).

6.1.1 Calcifications

Calcifications are the most common finding in mammography. Radiologists can find calcifications in lobules, ducts, interlobular tissue, vessels, skin, or soft tissues. As seen in Figure 6.1, BI-RADS defines that distribution (Figure 6.2) and morphology descriptions should be included in the medical report.

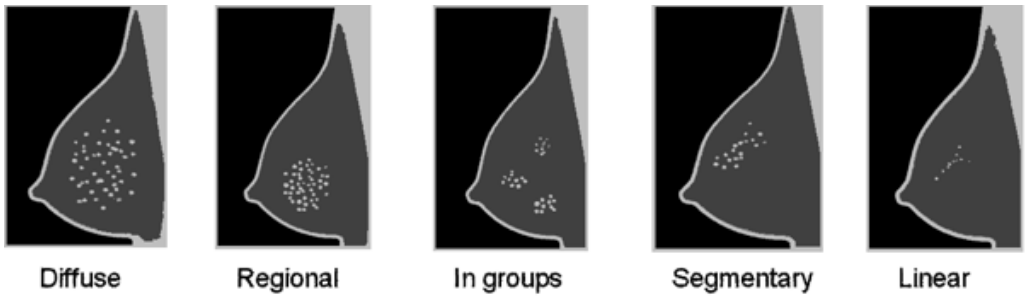


Figure 6.2: BI-RADS calcification distribution (Campos et al., 2013).

[1] A study on the use of ultrasound morphological features for mass classification is described in (Domingues et al., 2012a). Another publication related with this chapter is the segmentation method presented in (Cardoso et al., 2014).

In (Podsiadly-Marczykowska et al., 2009), calcifications were classified into three BI-RADS morphological categories, namely: punctuate (round), pleomorphic (calcifications of varying shapes and sizes) and linear. 27 shape descriptors were constructed and the morphology of a cluster was determined as the mean values of shape descriptors for single micro-calcifications. The most discriminative features are summarized in Table 6.1. Classification of Digital Database for Screening Mammography (DDSM) individual calcifications using SVMs obtained the following accuracies: 100% for linear versus round; 98% for linear versus pleomorphic; and 94% for pleomorphic versus round. Differentiation of clustered calcifications into linear or pleomorphic obtained an accuracy of 89%.

A content-based mammogram retrieval system is presented in (Wei et al., 2011). Four features were extracted to describe calcification types and another four to describe their distribution (Table 6.1). Diffuseness and the number of calcified spots were found to describe major calcification characteristics; dispersion, brightness and contrast are intermediate features; and spot size and shape describe the fine detail. Experimental results using the DDSM database showed that retrieval performance was improved when retrieving mammograms with similar mass and calcification lesions, respectively.

Another retrieval work is described in (Jing et al., 2012). The set of nine features selected by a sequential forward procedure were: number of calcifications in the cluster, cluster density, mean calcification size in the cluster, cluster eccentricity, standard deviation (*std*) of the distance from individual calcifications to the centre of the cluster, maximum of the mean intensity of the calcifications, mean of the average intensity in an 11×11 window around each calcification, *std* of the contrast of the calcifications and *std* of the fourth order central moment of the calcifications (Table 6.1). The proposed retrieval-driven approach was tested on a library of mammogram images from 1006 cases (646 benign and 360 malignant) which include DDSM images and images from a private database and obtained an area under the Receiver Operating Characteristic (ROC) curve of 0.78.

Wang and Yang (2013) investigate how to extract descriptive features for clustered calcifications such that they are robust for discriminating between malignant and benign lesions in the presence of detection inaccuracy. To this end, a Spatial Density Function (SDF) is proposed to characterize the spatial distribution of the calcifications in a cluster. The approach is demonstrated on a set of commonly used features (Table 6.1). Results using a private database and DDSM show that the SDF features are robust to variations in calcification detection while achieving better class separation.

6.1.2 Masses

A nodule is a space occupying lesion seen in two different projections. As seen in Figure 6.1, BI-RADS defines that margins (Figure 6.3), shape (Figure 6.4) and density (high, equal, low or fat-containing) descriptions should be included in the medical report.

The aim of the study in (Rangayyan et al., 2000) was to look at the least common shapes of masses such as circumscribed malignant tumours and spiculated benign masses which are difficult

Table 6.1: Some of the features used in the literature for calcification characterization.

Feature	Acronym
Area of the cluster region (Jing et al., 2012)	Acc
Binary Invariant Moments (Podsiadly-Marczykowska et al., 2009)	bIM
Brightness (Wei et al., 2011)	Br
Cluster compactness (Jing et al., 2012; Wang and Yang, 2013)	cCom
Cluster density (Jing et al., 2012; Wang and Yang, 2013)	cDen
Compactness (Jing et al., 2012)	Cm
Contrast (Jing et al., 2012; Wei et al., 2011)	Ct
Diffuseness (Wei et al., 2011)	Df
Dispersion (Wei et al., 2011)	Ds
Eccentricity (Jing et al., 2012; Wang and Yang, 2013)	Ecc
Effective Radius (Podsiadly-Marczykowska et al., 2009)	eR
Effective volume (Jing et al., 2012)	eV
Ellipticity (Podsiadly-Marczykowska et al., 2009)	Ell
Fourth order central moment (Jing et al., 2012)	M4
Inner Compactness (Podsiadly-Marczykowska et al., 2009)	IC
Intensity (Jing et al., 2012)	Int
Major axis (Podsiadly-Marczykowska et al., 2009)	Ma
Mean distance from cluster centroid (Jing et al., 2012)	dCC
Mean intensity of a window centred at the calcification (Jing et al., 2012)	Iwin
Number of Calcified Spots (Wei et al., 2011; Jing et al., 2012)	NrC
Roughness (Podsiadly-Marczykowska et al., 2009)	Rou
Roughness (Wang and Yang, 2013)	Rgh
Scatterness (Wang and Yang, 2013)	Sct
Shape irregularity (Jing et al., 2012)	Irr
Shape Moment (Podsiadly-Marczykowska et al., 2009)	ShM
Spot Shape (Wei et al., 2011)	SSh
Spot Size (Wei et al., 2011)	SSz
Std of distance from cluster centroid (Jing et al., 2012)	dCCstd

Figure 6.3: BI-RADS mass margins (Campos et al., 2013).

Figure 6.4: BI-RADS mass shapes (Campos et al., 2013).

to correctly classify using common shape analysis methods. These proposed methods of shape analysis treated the object's boundary in terms of local details. A boundary segmentation method was used to separate major portions of the boundary and to label them as concave or convex segments. To analyse the shape information localized in each segment, features were computed through an iterative procedure for polygonal modelling of mass boundaries. Features were based on the concavity fraction of a mass boundary and the degree of narrowness of spicules as char-

acterized by a spiculation index. 91% accuracy was attained for benign/malignant discrimination and 82% for circumscribed versus spiculated using the Mammographic Image Analysis Society (MIAS) digital mammogram database.

Sampat et al. (2005a) classified lesions into: round, oval, lobulated, or irregular. For this purpose, a new set of features using the Beamlet transform was developed. A k-Nearest Neighbours (kNN) classifier and images from DDSM were used. The method was tested on a set of 25 images of each type and obtained a classification accuracy of 78% for classifying masses as oval or round and an accuracy of 72% for classifying masses as lobulated or round.

Cheikhrouhou et al. (2008) presented new mass descriptors dedicated to differentiate between different mass shapes in mammography. A geometrical feature, perimeter, and three morphological features were used. These features were: contour derivative variation, skeleton end points and spiculation. Their performance was evaluated on DDSM using SVMs with a Gaussian kernel as classifier. The accuracy was 93% in the two class case (malignant and benign) and 85.7% in the four class model (BI-RADS I, II, III and IV). The work presented in (Cheikhrouhou et al., 2008) differs from the classification technique of Chapter 7 in two main ways: Cheikhrouhou et al. (2008) classify mass regions only, while in (Domingues and Cardoso, 2013b) a classification of the full mammogram image is made; in addition, the work of Cheikhrouhou et al. (2008) does not use the fact that the BI-RADS scale is ordinal, while this information is used in (Domingues and Cardoso, 2013b) to develop a better classifier.

Tao et al. (2008) present a clinically guided technical method for content-based categorization of mammographic masses. Four experiments were made: classification into (1) Regular, Lobular or Irregular, (2) Regular or Irregular, (3) Circumscribed, Microlobulated, Indistinct or Spiculated, and (4) Circumscribed, Indistinct or Spiculated. Shape, acutance histogram and texton features were extracted and experiments were conducted both on a private database and on DSSM. In the experiments that categorize lesion shape, a precision of 70% with three classes (experiment 1) and 87.4% with two classes (experiment 2) was obtained. In the experiments that categorize the margins, precisions of 69.4% and 74.7% were obtained for the use of four (experiment 3) and three (experiment 4) classes, respectively.

Rojas-Dominguez and Nandi (2009) present four new features designed to be insensitive to the exact shape of the contour of the masses. Two of the features measure the degree of spiculation of a mass and its likelihood of being spiculated (while one is a measure of the relative gradient orientation of pixels that correspond to possible spicules, the other is based on a comparison of mutual information measures between selected components of the mammographic images). The last two features measure the local fuzziness of the mass margins based on points defined automatically. The features were tested for characterization (circumscribed/spiculated) and diagnosis (benign/malignant) using mini-MIAS and DDSM. In the characterization experiments the features produced a result of approximately 89% correct classification. In the diagnosis experiments, the performance achieved was approximately 81% correct classification.

Two new shape measures for quantifying the degree of convexity are proposed by Rosin (2009). The first is based on convexification, while the second is based on contained lines. The

shape measures were applied to a set of 54 masses from mammograms, combining images from MIAS and a private database. Classification was performed using kNN with the Mahalanobis distance. A correct classification of 94.4% was achieved on circumscribed/spiculated discrimination, 74.1% for benign/malignant discrimination and 68.4% for circumscribed benign/circumscribed malignant/spiculated benign/spiculated malignant discrimination.

The content-based mammogram retrieval system presented in (Wei et al., 2011) and already mentioned in Section 6.1.1 also deals with masses. Three mass features are used to describe shape (Zernike moments), margin (sharpness degree), and density (density degree). As for all the remaining works mentioned in this section, the features used for mass characterization are summarized in Table 6.2.

A novel descriptor invariant to rotation was developed to deal with round-shaped objects, such as masses, by Moura and Lopez (2013). The proposed descriptor, Histograms of Gradient Divergence (HGD), was compared with 11 conventional image descriptors using cases from DDSM and BCDR. Overall, HGD scored best (or comparable to best) on both DDSM and BCDR when classifying masses as benign or malignant.

Ribeiro (2013) used texture and geometric descriptors to characterize masses from DDSM and a private database. The selected classifier was an ANN. The algorithm for the distinction between ROIs with and without masses achieved an accuracy of 83.75%. Classification into round masses or architectural distortion achieved an accuracy of 74.02%. Distinction between the 5 BI-RADS mass margins classes (see Figure 6.3) attained an accuracy of 71.00% for the private database and 39.00% for DDSM. Grouping the mass margins classes into two sets (circumscribed and partially obscured versus microlobulated, ill-delimited and spiculated), increased the accuracy to 82.96% for the private database and 82.38% for DDSM. Focusing only on circumscribed and spiculated masses, the algorithm produced correct classifications in 87.38% of the cases.

Vadivel and Surendiran (2013) presented new geometric shape and margin features for classifying mammogram mass lesions into: round, oval, lobular and irregular. Experiments were conducted on mammogram images from DDSM and classified using a C5.0 decision tree classifier. Obtained accuracies were: 87.76% (irregular, lobular, oval or round), 100% (oval versus round), and 95.45% (lobulated versus round). It is claimed that this approach is twice as effective as the Beamlet based features in (Sampat et al., 2005a).

In Tan et al. (2014a), 181 image features based on mass shape, spiculation, contrast, presence of fat or calcifications, texture, isodensity, and other morphological features were first calculated. Then, a sequential forward floating selection-based feature selection method was used to select relevant features. Their performance was analysed using an SVM model trained for the classification task. The used database consists of 1 200 ROIs (600 malignant masses and 600 benign) randomly selected from a private database and DDSM. The obtained area under the ROC curve was 0.805 ± 0.012. The results also showed that the most frequently selected features were those related to mass shape, isodensity, and presence of fat, which are consistent with the image features frequently used by radiologists in the clinical environment for mass classification.

Table 6.2: Some of the features used in the literature for mass characterization.

Feature	Acronym
Acutance histogram (Tao et al., 2008)	AcH
Area (Ribeiro, 2013; Vadivel and Surendiran, 2013)	A
Beamlet (Sampat et al., 2005a)	Beam
Circularity (Vadivel and Surendiran, 2013)	Circ
Compactness (Rangayyan et al., 2000; Vadivel and Surendiran, 2013)	Com
Concavity fraction (Rangayyan et al., 2000)	fCC
Contained lines (Rosin, 2009)	Cl
Contour Derivative Variation (Cheikhrouhou et al., 2008)	CDV
Convexification (Rosin, 2009)	Cvf
Convexity fraction (Rangayyan et al., 2000)	fCV
Curvature Scale Space (Tao et al., 2008)	CSSD
Curvelets (Moura and Lopez, 2013)	Curv
Dispersion (Vadivel and Surendiran, 2013)	Dp
Eccentricity (Vadivel and Surendiran, 2013)	ECT
Elongatedness (Vadivel and Surendiran, 2013)	En
Entropy (Vadivel and Surendiran, 2013)	Entpy
Equivalent diameter (Vadivel and Surendiran, 2013)	Eqd
Euler number (Vadivel and Surendiran, 2013)	EULN
Fourier (Tao et al., 2008)	NFD
Fuzziness of mass margins (Rojas-Dominguez and Nandi, 2009)	FZk
Gabor filter banks (Moura and Lopez, 2013)	Gab
Grey-level difference matrix (Moura and Lopez, 2013)	GLDM
Grey-level run length (Moura and Lopez, 2013)	GLRL
Histograms of Gradient Divergence (Moura and Lopez, 2013)	HGD
Histograms of oriented gradient (Moura and Lopez, 2013)	HOG
Mass edge *std* (Vadivel and Surendiran, 2013)	Esd
Mass *std* (Vadivel and Surendiran, 2013)	SD
Maximum radius (Vadivel and Surendiran, 2013)	Rmax
Minimum radius (Vadivel and Surendiran, 2013)	Rmin
Perimeter (Cheikhrouhou et al., 2008; Vadivel and Surendiran, 2013)	Per
Radial to tangential signature (Rojas-Dominguez and Nandi, 2009)	SpSI
Relative gradient orientation spiculation (Rojas-Dominguez and Nandi, 2009)	SpGO
Shape Index (Vadivel and Surendiran, 2013)	ShI
Sharpness (Wei et al., 2011)	Sh
Skeleton end points (Cheikhrouhou et al., 2008)	SEP
Spiculation Index (Rangayyan et al., 2000)	SpI
Spiculation (Cheikhrouhou et al., 2008)	Sp
Texton (Tao et al., 2008)	Txo
Thinness ratio (Vadivel and Surendiran, 2013)	Thi
Wavelets (Moura and Lopez, 2013)	Wav
Zernike Moments (Wei et al., 2011; Moura and Lopez, 2013)	Zm

6.2 Feature extraction

As seen in the previous section, there is a long list of commonly used features both for calcification and masses. In order to evaluate their relevance for the INbreast database, some of these features were tested. The next sections list the features that were extracted in this work from the ground truth information marking the presence of calcifications and masses.

6.2.1 Calcifications

In Section 5.2 a technique was presented for finding calcification centres. This is, however, a detection-only approach and does not include any segmentation algorithm. As a consequence, calcification contours are not available, meaning all the existing calcification shape features were excluded. From Table 6.1, 16 types of characteristics were extracted: Acc, Br, cCom, cDen, dCC, dCCstd, Df, Ds, Ecc, Int, Iwin, NrC, Rgh, Rou, Sct and ShM. Normalized versions of dCC,

dCCstd, Df were also computed by dividing by the length of the image diagonal. A normalized version of Acc was obtained dividing by the area of the image. Note that not all of the above characteristics correspond to a single feature, e.g., Ds is of size 15. In this way, a 53 dimensional feature vector for each calcification candidate is obtained.

6.2.2 Masses

Contrary to the calcification detection algorithm, the mass detection system (presented in Section 5.3) does have a mass segmentation step. Image segmentation, although it has been extensively studied (Suri et al., 2002), is still an open problem. Shortest path algorithms are often used to tackle this problem. There are, however, applications where the starting and ending positions of the shortest path need to be constrained to be the same, defining a closed contour to enclose a previously detected seed. Mass segmentation in mammograms is an example where this constraint holds.

Usually the closed contour computation is addressed by transforming the image into polar coordinates, where the closed contour is transformed into an open contour between two opposite margins. In this transformation resolution may be lost, discontinuities introduced and extra computational power and time required. Additionally, the distance between neighbouring pixels has to be defined in the transformed space. In addition to being less intuitive, it will vary with the angular and radial resolution chosen for the polar image.

In this work the closed contour is computed in the original coordinate space (Cardoso et al., 2014) (this method will be referred to as "CaPTOR"). An appropriate directed acyclic graph is first defined. Since small paths collapsing in the seed point are inherently favoured, the cost of the edges is modulated to counterbalance this bias. The algorithm was shown to be fast and reliable and suffers no loss in resolution. A study of this segmentation technique is shown in section 6.3.3.

Concerning the features used for mass description, 32 types of characteristics from Table 6.2 were extracted: A, CDV, Circ, Cl, Com, Curv, Cvf, DD, Dp, ECT, En, Entpy, Eqd, Esd, EULN, Gab, GLDM, GLRL, HGD, HOG, Per, Rmax, Rmin, SD, SEP, Sh, ShI, Sp, Thi, Txt, Wav and Zm. Other features not present in Table 6.2 but commonly used in the computer vision field were also extracted using the Matlab *regionprops* function. These include:

- convex area (Aconv);
- distance between the centroid and the weighted centroid (dCwC);
- extent (Ext);
- filled area (Afill);
- major axis length (ALmaj);
- maximum intensity (IntMax);
- mean intensity (IntAvg);

- minimum intensity (IntMin).
- minor axis length (ALmin);
- orientation (Ori);
- solidity (Sol);

Normalized versions of ALmaj, ALmin, dCwC, and Eqd were also extracted by dividing by the length of the image diagonal. Aconv and Afill normalized versions were obtained by dividing by the image area. A normalized version of Per was obtained by dividing by the area of the mass. Finally, a normalized version of Ori, whose original units are in degrees ranging from -90 to 90 degrees, was calculated by adding 90 and dividing by 180.

Descriptors Curv (size 2 761), Gab (size 1 121), GLDM (size 121), GLRL (size 45), HGD (size 10 801), Wav (size 1 569), and Zm (size 16) were summarized into single features by calculating mean average values. HOG with blocks of 3×3 of original size 433 and HOG with blocks of 5×5 of original size 1201 are summarized in the same way, generating features HOG_3 and HOG_5 respectively. In this way, a final vector of size 56 for each mammogram image was obtained.

6.3 Results

In the experiments made in this chapter it is assumed that: orientation homogenization has been performed, the breast region has been detected, and images with BI-RADS class 1 have been screened out. Moreover, BI-RADS classes 5 and 6 were merged. Unless stated otherwise, features are extracted using the Ground Truth (GT) detections. Results using the automatic detection methods presented in Chapter 5 will be provided in Chapter 8.

6.3.1 Feature selection

In order to test which features are informative, three different metrics were used, namely the Pearson correlation, the distance correlation and the Maximal information coefficient.

The Pearson correlation (*corr*) is the most commonly used measure of the three and quantifies the linear dependence between two variables. *corr* can assume values between -1 and $+1$, inclusive, where $+1$ is total positive correlation, 0 is no correlation, and -1 is total negative correlation.

The distance correlation (*dcorr*) characterizes independence: it is zero if and only if the vectors are independent (Szekely and Rizzo, 2009). Comparing with the Pearson correlation, *dcorr* measures not only linear associations, but all types of dependence relations. The distance correlation satisfies $0 \leq dcorr \leq 1$.

The Maximal information coefficient (MIC) is based on the intuitive idea that if a relationship exists between two variables, then a grid can be drawn on the scatter-plot of the two variables that partitions the data to encapsulate that relationship (Reshef et al., 2011). MIC is symmetric $\text{MIC}(X,Y) = \text{MIC}(Y,X)$ and its values range from 0 to 1. MIC assigns values that tend to 1 for

a large class of noiseless functional relationships and assigns values that tend to 0 to statistically independent variables.

6.3.2 Calcification characterization

The experiments presented in this section were made with the INbreast images containing calcifications only. In order to test if the features are informative, the correlation between each feature vector and the corresponding class was computed. The hypothesis of no correlation against the alternative that there is a non-zero correlation was tested. Seven features had p-values smaller than 0.1 and were thus kept. The top seven features retrieved by *dcorr* and MIC were also computed. Results are summarized in Table 6.3 and respective box-plots are presented in Figure 6.5.

Table 6.3: Selected calcification features.

Features		*corr*	*dcorr*	MIC
Symbol	Description			
Ds_4^0	Zernike moment of order 4 and repetition 0	0.212	0.231	
Ds_3^{+3}	Zernike moment of order 3 and repetition $+3$	0.201	0.223	0.068
Ecc_{sdf}	Eccentricity extracted from SDF	0.146	0.164	
Ds_4^{-4}	Zernike moment of order 4 and repetition -4	0.139	0.169	0.062
Ds_4^{+2}	Zernike moment of order 4 and repetition $+2$	0.123		
Ds_3^{-1}	Zernike moment of order 3 and repetition -1	0.114		
Sct_{sdf}	Scatterness extracted from SDF	0.111		
$Iwin_{min}$	Minimum of the mean intensities of the calcifications		0.175	
Int_{std}	Intensity *std*		0.141	
$Iwin_{std}$	*Std* of the mean intensities of the calcifications		0.141	
Int_{max}	Intensity maximum			0.088
ShM_6	Shape Moment of order 6			0.070
Int_{min}	Intensity minimum			0.069
ShM_4	Shape Moment of order 4			0.067
Ds_4^{+4}	Zernike moment of order 4 and repetition $+4$			0.065

Features Ds_3^{+3} and Ds_4^{-4} are on the top seven of all tested methods. Ds_4^0 and Ecc_{sdf} are selected both by correlation and *dcorr*. All the other features are selected by only one of the techniques.

6.3.3 Mass characterization

This section is divided into two parts. In the first part (section 6.3.3.1), the segmentation method is evaluated. The second part (section 6.3.3.2) presents feature visualization and selection results.

6.3.3.1 Mass segmentation

All 116 masses from the INbreast database were used in this subsection. The rectangular Region Of Interest (ROI) was generated from the Bounding Box (BB) of each mass, by expanding the BB by 20%.

In order to assess the behaviour of the CaPTOR segmentation in comparison with existing state of the art methods, other techniques were tested on this database. To act as a baseline technique, an ellipse (this method will be referred to as "Ellipse") was generated centred in the ROI and with axis diameters corresponding to $1/1.2$ the ROI length (or width) to compensate for the 20% of the

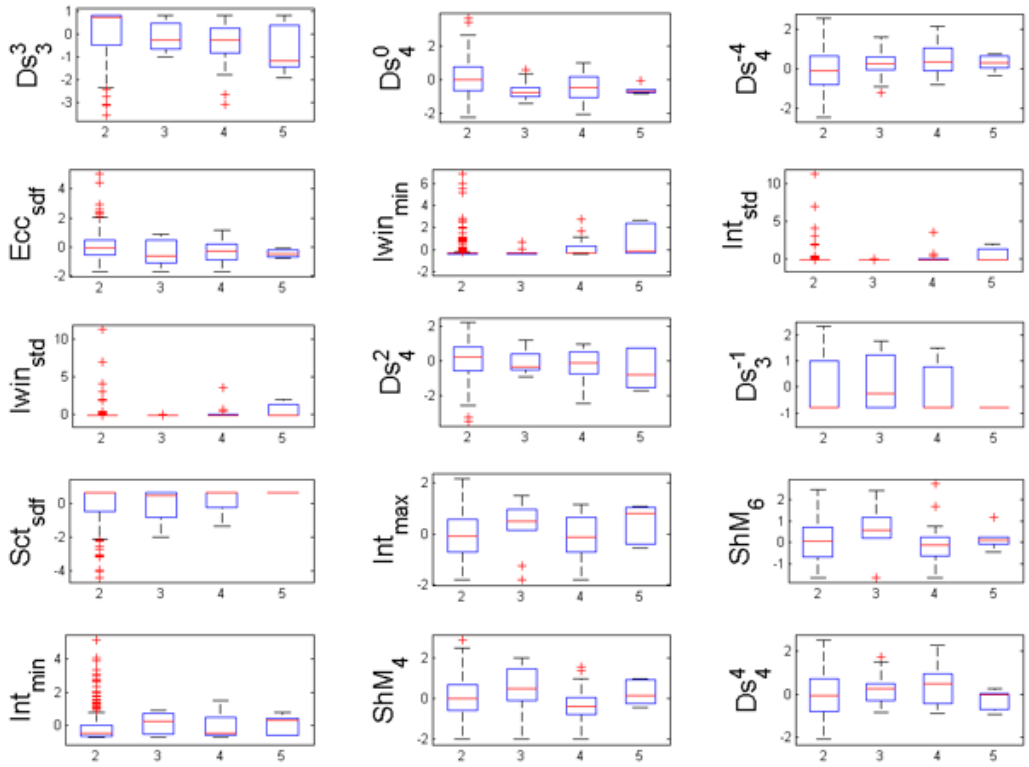


Figure 6.5: Box-plots of calcification features. BI-RADS assessment is represented in the horizontal axis.

mass surrounding region included in the ROI. Note that this baseline method uses knowledge not available in practice but it helps us attain a useful reference result.

The second method is based on the active contours idea, implementing the basic snake segmentation contour (this method will be referred to as "Snake"), as introduced by Kass et al. (1988), including a "balloon force" parametrization, which is not part of the original methodology[1]. The method requires an initial step for the positioning of an approximate initial contour. In the present tests, the previously generated ellipse was used as initialization.

The third comparison method is the graph-cut method with a star shape prior (Veksler, 2008) (this method will be referred to as "Star"). The implementation was provided by Gulshan et al. (2010)[2]. The "ESC" segmentation algorithm was selected and the default parameter values were used. Initialization was manually made by generating two strokes inside the mass region and two strokes outside the mass.

Finally, a local convergence filter (LCF) with a sliding band filter (this method will be referred to as "SBF") was tested[3]. LCF detects the locations in the image where the image gradient converges. This makes it well-suited for finding convex shapes. The SBF defines a support region

[1]Code available at `http://www.mathworks.com/matlabcentral/fileexchange/28149-snake-active-contour`.

[2]Code available at `http://www.robots.ox.ac.uk/~vgg/research/iseg/`.

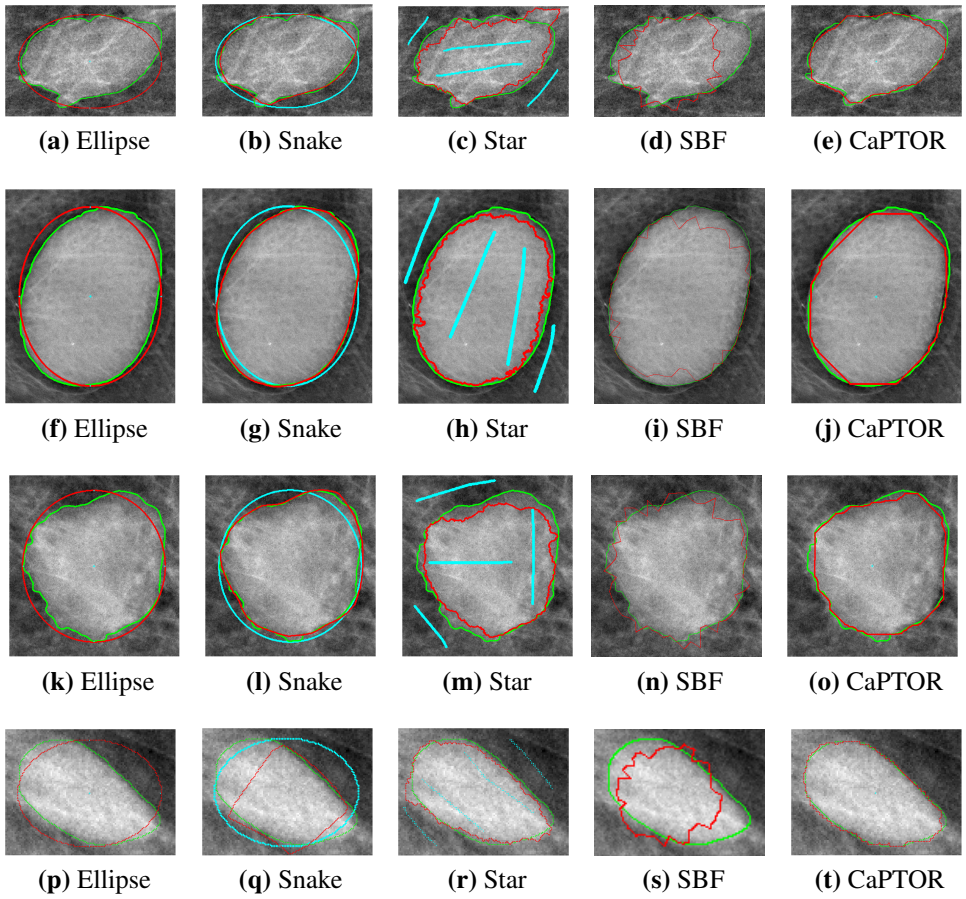
[3]Code available at `http://paginas.fe.up.pt/~quelhas/LCFs/`.

formed by a band of fixed width, with varying radii in each direction to allow for the maximization of the convergence index at each point. For more details on this technique, please refer to (Esteves et al., 2012).

Results of these techniques are shown in Table 6.4. Example illustrations of some segmentations are shown in Figure 6.6.

Method	AOM	CM	AD	AMED	HD
Ellipse	0.754 (0.112)	0.830 (0.078)	0.036 (0.016)	0.038 (0.018)	0.119 (0.056)
Snake	0.767 (0.118)	0.839 (0.084)	0.032 (0.017)	0.035 (0.019)	0.127 (0.056)
Star	0.753 (0.093)	0.827 (0.072)	0.029 (0.015)	0.032 (0.017)	0.118 (0.048)
SBF	0.724 (0.105)	0.805 (0.078)	0.034 (0.014)	0.037 (0.015)	0.131 (0.047)
CaPTOR	0.792 (0.104)	0.858 (0.072)	0.024 (0.012)	0.028 (0.015)	0.105 (0.045)

Table 6.4: Mass segmentation: results for different segmentation methods. Results are presented in the format: mean (*std*).



(a) Ellipse **(b)** Snake **(c)** Star **(d)** SBF **(e)** CaPTOR

(f) Ellipse **(g)** Snake **(h)** Star **(i)** SBF **(j)** CaPTOR

(k) Ellipse **(l)** Snake **(m)** Star **(n)** SBF **(o)** CaPTOR

(p) Ellipse **(q)** Snake **(r)** Star **(s)** SBF **(t)** CaPTOR

Figure 6.6: Example illustrations of some mass segmentations using the different methods. The GT contour is shown in green and the automatic contour is shown in red. Initialization is shown in cyan. For a better contour illustration, the inter-image proportion was not kept constant.

According to the Region metrics (AOM and CM), SBF is significantly worse than all the other methods, whereas the CaPTOR method is significantly better than every other method. For

both AD and HD, the CaPTOR still shows a significant improvement over all the other tested approaches. The only exception is when comparing CaPTOR with the Star technique using AMED, in which case no significant difference was found.

Some other works have computed mass contours based on the shortest path algorithm in polar coordinates. These are summarized in Table 6.5. The closed shortest path in polar coordinates

Paper	AOM	CM	AD
Timp and Karssemeijer (2004)	0.69	-	-
Dominguez and Nandi (2009)	0.72	-	-
Elter et al. (2010)	-	0.70	-
Song et al. (2010)	0.73	-	-
Xu et al. (2010)	-	-	2.72
Qi et al. (2012)	0.70	-	-
Yu et al. (2012)	0.65	-	-
Zhang et al. (2012)	0.70	-	-

Table 6.5: Mass segmentation published results.

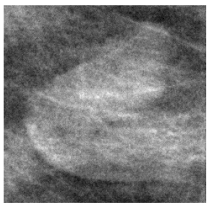
was also tested in INbreast, obtaining an AOM of 0.74. Since, in general, the databases are not the same, a direct comparison cannot be made. Moreover, most of the papers use region-based evaluation metrics and not contour-based metrics. However, the CaPTOR method appears to have a competitive accuracy when compared to previous implementations. Moreover, in INbreast, CaPTOR attains a better performance than the traditional polar coordinates approach, probably due to the impact of the angular and radial resolution of the final result in the traditional method. Figure 6.7 illustrates this behaviour in one of the masses.

6.3.3.2 Mass features

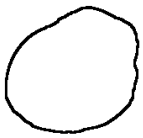
The experiments presented in this section were made with the INbreast images containing masses only. As in the calcifications case, in order to test if the features were informative, the correlation between each feature vector and the corresponding BI-RADS class was computed. The hypothesis of no correlation against the alternative that there is a non-zero correlation was tested. Nine features had p-values smaller than 0.1 and were thus kept. The top nine selected features by correlation, *dcorr* and MIC are summarized in Table 6.6 and box-plots are shown in Figure 6.8.

Table 6.6: Selected mass features.

Features		*corr*	*dcorr*	MIC
Symbol	Description			
Sol	Solidity	0.660	0.675	0.437
Thi	Thinness ratio	0.631	0.640	0.348
Com	Compactness	0.622	0.642	0.432
Cvf	Convexification	0.576	0.658	
Cl	Contained lines	0.553	0.664	
SEP	Skeleton end points	0.531	0.546	0.346
ShI	Shape Index	0.490	0.509	0.352
Ext	Extent	0.463	0.462	0.292
CC_2	$CC_2 = \sqrt{\frac{Rmin}{Rmax}}$	0.388	0.398	
ALmaj	Major axis length			0.301
$ALmaj_{norm}$	Normalized major axis length			0.301
ECT	Eccentricity			0.279



(a) Original ROI with mass.

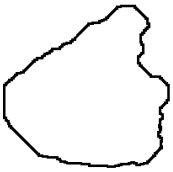


(b) Output by the traditional polar transformation with 1°/pixel.

(c) Output by the traditional polar transformation with 0.16°/pixel.

(d) Reference segmentation.

(e) Output by CaPTOR with 1°/pixel.



(f) Output by CaPTOR with 0.16°/pixel.

Figure 6.7: Impact of angular resolution in the final segmentation. The radial resolution was kept constant.

Features Sol, Com, Thi, SEP, ShI and Ext were selected by the three techniques. Cvf, Cl and CC_2 were selected by both correlation and *dcorr*. ALmaj, ALmaj_{max} and ECT were chosen by MIC only. It is clear in this example that the selected metrics do not take into consideration the interaction between features.

It is also interesting to study the influence of the automatic mass segmentation method in the selected features. The box-plots of the 12 features on Table 6.6 extracted from the CaPTOR contours are shown in Figure 6.9. In this dataset, only five features pass the correlation test. They are: Cl, Cvf, Ext, SEP and Sol. The top five features selected by *dcorr* coincide with the ones chosen by *corr*. MIC selects: ALmaj, ALmaj_{norm}, Ext, SEP and Sol, having in common with the other methods only three of the top five features.

6.4 Conclusions

In this chapter a literature review of methods that classify calcifications into BI-RADS distribution and morphology categories and masses into BI-RADS margins, shape and density categories was performed giving special emphasis to the automatic features used for classification. A lack of feature selection methods described in the literature for selection of continuous features for ordinal

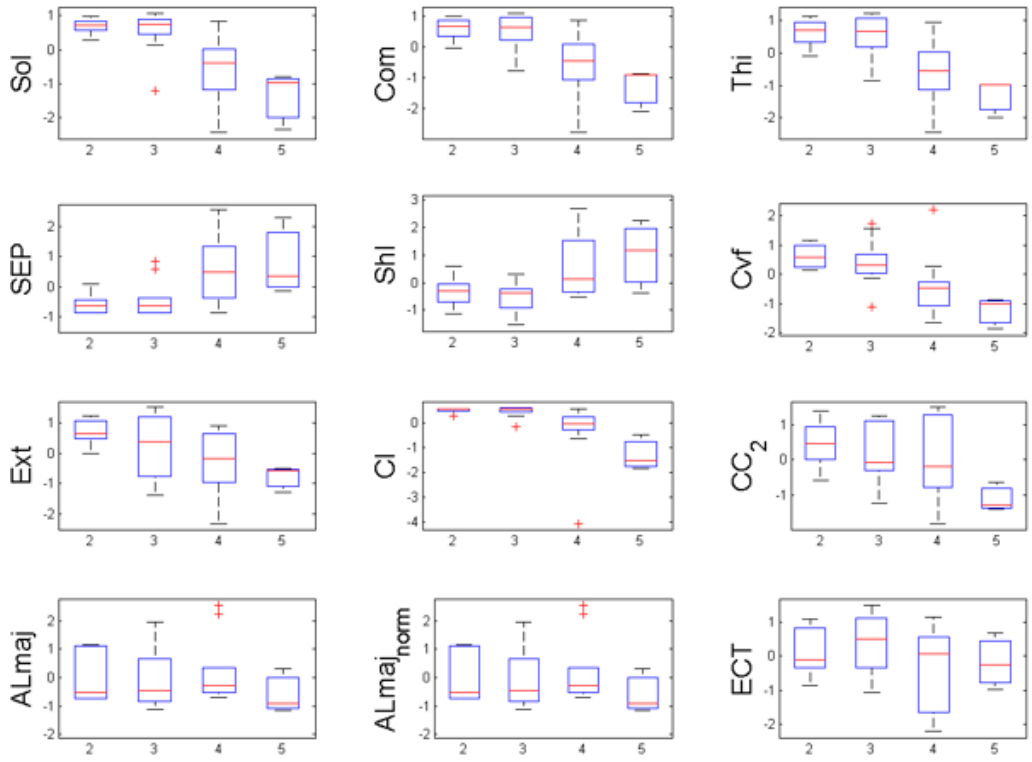


Figure 6.8: Box-plots of mass features extracted from GT contours. BI-RADS assessment is represented in the horizontal axis.

classification was observed. In this way, a large portion of the existing features was evaluated using INbreast database by using Pearson correlation, distance correlation and MIC.

A subset of seven calcification features were shown to have a significant correlation with the BI-RADS final assessment. For masses, a final vector of nine features was built. The automatic mass segmentation method was also evaluated. It was concluded that from the previous nine features, only five were significant when extracted from the automatic contours. The features here studied will be used as input to the BI-RADS classifier proposed in the next chapter.

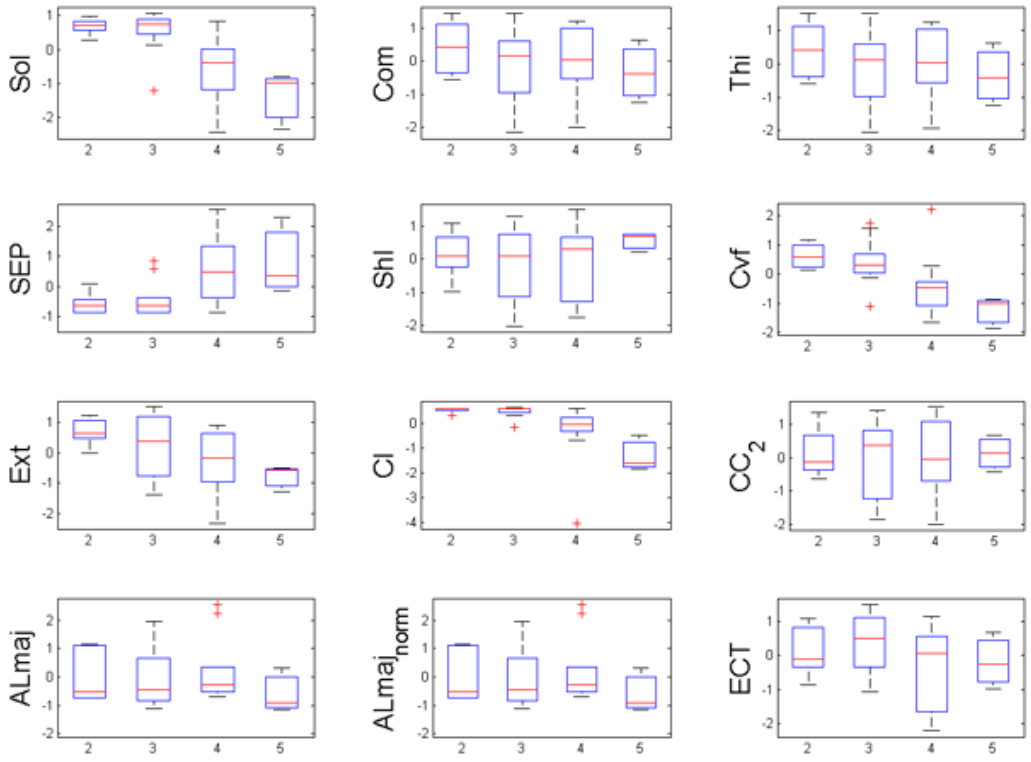


Figure 6.9: Box-plots of mass features extracted from automatic contours. BI-RADS assessment is represented in the horizontal axis.

Chapter 7

BI-RADS classification

When radiologists examine mammograms, they look for specific abnormalities. As already seen, breasts can have: only mass(es), only calcifications(s) or both. Based on the level of suspicion, lesions can be placed into one of seven BI-RADS categories (Table 2.3): 0 when the exam is not conclusive, 1 for no findings, 2 for benign findings, 3 for probably benign findings, 4 for suspicious findings, 5 when there is a large probability of malignancy, and 6 for proven cancer.

From the perspective of a supervised learning setting, the prediction of the malignancy of a case can be addressed as a multi-class classification problem, where there is a natural order among the classes: this corresponds to an ordinal data classification problem, with higher Breast Imaging Reporting And Data System (BI-RADS) values corresponding to higher probabilities of malignancy. However, when more than one finding is present in the mammogram, the overall BI-RADS in the medical report corresponds to the finding with the highest BI-RADS. Therefore the knowledge about the training examples is neither complete nor totally incomplete.

In this chapter[1] the concept of Max-Ordinal Learning (MOL) is introduced and studied. The problem is formalized in Section 7.1, next the related state of the art is summarized in Section 7.2. This state of the art focuses on learning methodologies since breast cancer classification methods were already summarized in Section 6.1. The proposed methodologies to deal with MOL problems are described in Section 7.3 and evaluated in Section 7.4.

7.1 Problem statement

In statistical pattern recognition, it is usually assumed that a training set of labelled patterns is available where each pair $\{\mathbf{x}_i, y_i\} \in \Re^d \times \mathscr{Y}$ has been generated independently from an unknown distribution. The goal is to induce a classifier, i.e., a function from patterns to labels $f : \Re^d \rightarrow \mathscr{Y}$. This work focuses on the ordinal case of $\mathscr{Y} = \{y_1, \cdots, y_K\}$, where $y_1 \prec \cdots \prec y_K$ and $\prec$ is a linear order relation in $\mathscr{Y}$.

[1] Publications related with this chapter include (Domingues and Cardoso, 2013b, 2012; Cardoso et al., 2012; Domingues and Cardoso, 2011; Cardoso and Domingues, 2011).

MOL generalizes this problem by making weaker assumptions about the labelling information. The labelled patterns in the training set can be one of three types that are organized into three different subsets, as follows:

1. $\mathscr{S}_1 = \{\mathbf{x}_i, y_i = f(\mathbf{x}_i)\}_{i=1}^{N_1}$, where $\mathbf{x}_i \in \Re^{d_1}$, $y_i \in \mathscr{Y}$ and $f(.)$ is unknown.
2. $\mathscr{S}_2 = \{\mathbf{z}_i, y_i = g(\mathbf{z}_i)\}_{i=1}^{N_2}$, where $\mathbf{z}_i \in \Re^{d_2}$, $y_i \in \mathscr{Y}$ and $g(.)$ is unknown.
3. $\mathscr{S}_{12} = \{\mathbf{x}_i, \mathbf{z}_i, y_i\}_{i=1}^{N_{12}}$, where $\mathbf{x}_i \in \Re^{d_1}$ and $\mathbf{z}_i \in \Re^{d_2}$, $y_i \in \mathscr{Y}$. It is known that $y_i = \max(f(\mathbf{x}_i), g(\mathbf{z}_i))$ but both $f(\mathbf{x}_i)$ and $g(\mathbf{z}_i)$ are unobserved.

When both views $\mathbf{x}$ and $\mathbf{z}$ are present, the individual classification of each view is unobserved and only the maximum of both is known. An illustration is given in Figure 7.1 (note that $\mathscr{S}_1$, $\mathscr{S}_2$ and $\mathscr{S}_{12}$ are disjoint sets). When $\mathscr{S}_{12}$ is empty, this results in the learning of two independent classifiers.

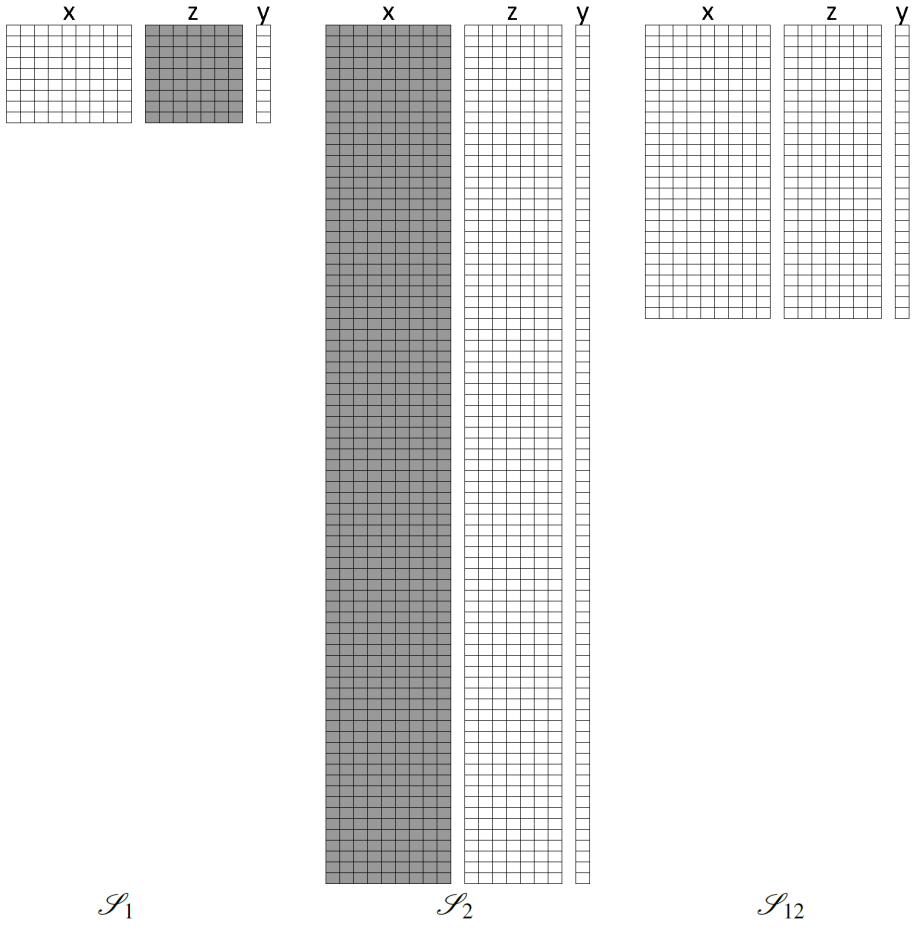


Figure 7.1: Training set illustration. White represents observed and grey represents not present features.

In the breast cancer application, y_i corresponds to a known classification in one of the BI-RADS ordinal classes present in the medical report. The subsets are as follows: 1) $\mathscr{S}_1$ corresponds to the cases where only a mass was detected in the mammogram; 2) $\mathscr{S}_2$ corresponds to

the cases where only calcifications were detected in the mammogram; 3) $\mathscr{S}_{12}$ corresponds to the cases where both mass and calcifications were detected in the mammogram. The function $f(\mathbf{x}_i)$ corresponds to the BI-RADS classification due to the presence of the mass only; and similarly for $g(\mathbf{z}_i)$. Remember that the medical report only includes the classification corresponding to the highest BI-RADS.

The learning problem can be formulated as seeking $f(.)$ and $g(.)$ that minimize the expected loss over the distribution of observations, for a pre-specified loss function. In general, the risk (expectation of the loss function) cannot be computed because the underlying distribution is unknown and functions $f(.)$ and $g(.)$ are selected based on the performance in the training set (empirical risk):

$$\{f^*, g^*\} = \arg\min_{f,g} \left(\sum_{\mathbf{x}_i \in \mathscr{S}_1} \mathscr{L}(f(\mathbf{x}_i), y_i) + \sum_{\mathbf{z}_i \in \mathscr{S}_2} \mathscr{L}(g(\mathbf{z}_i), y_i) + \sum_{\mathbf{x}_i, \mathbf{z}_i \in \mathscr{S}_{12}} \mathscr{L}(\max(f(\mathbf{x}_i), g(\mathbf{z}_i)), y_i) \right) \tag{7.1}$$

where $\mathscr{L}$ is a loss function. A typical loss function for ordinal data is the Mean Absolute Error (MAE). Note that this problem formulation is valid only for ordinal data, since the maximum is not defined for nominal classes.

It is interesting to mention two special cases. The first is the extension of this model to multiple views, M (instead of just two). Each observation only includes a subset of the views and the label corresponding to the maximum of the individual views (the individual label of each view included is unknown). Another setting of interest is when all observations are from the third type. This means that the two (or more) views are always observed and the label corresponds to the maximum of the unknown individual labels. These two special cases will be briefly addressed at the end of the description of each of the proposed techniques.

7.2 Related work

In 2001, Frank and Hall (2001) introduced a simple process which made it possible to explore information order in classification problems, using conventional binary classifiers. The problem is transformed from a K-class ordinal problem to $K-1$ binary class problems. The main advantage of this scheme is that any binary classifier can be used as the building block.

In 2007, Cardoso presented the data replication method (Cardoso and Costa, 2007), a Single Binary Classifier (SBC) approach to solve multi-class problems via binary classification relying on a single, standard binary classifier. SBC reductions can be obtained by embedding the original problem in a higher-dimensional space consisting of the original features, as well as one or more extension features. This embedding is implemented by replicating the training set points so that a copy of the original point is concatenated with each of the extension features' vectors. The binary labels of the replicated points are set to maintain a particular structure in the extended space. This construction results in an instance of an artificial binary problem, which is fed to a single binary

learning algorithm. As in (Frank and Hall, 2001), any binary classifier can be used as the building block.

Another SBC technique that solves the binary classification sub-tasks with point-wise regression is given in (Ruan et al., 2014). The method is named Cost-sensitive Ordinal Classification via Regression (COCR), and allows the specification of mis-ranking costs to further improve the ranking performance. This ability is exploited by deriving a corresponding cost for Expected Reciprocal Rank (ERR). The resulting ERR-tuned COCR boosts the benefits of the efficiency of using point-wise regression and the accuracy of top-rank prediction from the ERR criterion.

The Kernel Discriminant Learning Ordinal Regression (KDLOR) method (Sun et al., 2010a) was proposed in 2010. KDLOR is an adaptation of the conventional Linear Discriminant Analysis (LDA) method with a ranking constraint. The main goal is to find the optimal linear projection for classification (from which different classes can be well separated) while preserving the ordinal information of classes, i.e., the average projection of the samples from the higher rank classes should be larger than that of lower rank classes. The original LDA optimization problem is transformed and extended with a penalty term to account for the constraint in the projected means. To accommodate non-linear problems, the algorithm is modified to incorporate the kernel trick.

A shortcoming of the KDLOR algorithm is that only one dimension in the sample space is used, which loses some useful information in its orthogonal subspaces. In (Sun et al., 2014), a novel ordinal regression strategy that can extract multiple features from the original data space is proposed. It consists of two stages: firstly orthogonal feature vectors are extracted and then these projection vectors are combined to learn an ordinal regression rule.

Techniques for Semi-Supervised Learning (SSL) are common nowadays. Co-training and multi-view models are representative examples which assume that there are multiple, different learners trained on the same labelled data, and these learners agree on the unlabelled data. For example, in tri-training (Zhou and Li, 2005) the labelled data are split in three sets and a classifier is trained in each set. If two of them agree on the classification of an unlabelled point, the classification is used to "teach" the other classifier.

A Transductive Ordinal Regression (TOR) paradigm involving labelled and unlabelled data for learning ordinal decision functions was introduced in (Seah et al., 2012). A label swapping scheme for multiple ordinal class transduction was also proposed. Numerical results show that this transductive approach achieves significant accuracy improvements in terms of mean zero-one and absolute errors.

In learning methodologies with incomplete knowledge, perhaps the most similar methodology to the one addressed here is Multiple Instance Learning (MIL) (Andrews et al., 2003). The basic idea of MIL is that, during training, examples are presented in sets (often called "bags"), and labels are provided for the bags rather than for the individual instances. If a bag is labelled positive, it is assumed to contain at least one positive instance, otherwise the bag is labelled negative (Zhang et al., 2014b). Note that this paradigm is for binary settings only and that all the observations in the bag come from the same "view" (feature set). Felzenszwalb et al. (2010) use a latent variable formulation of the above called MIL SVM (Andrews et al., 2003) to train models using partially

labelled data. Once again, this formulation only applies to the binary case. Furthermore, the classification function is the maximum of linear functions (in our approaches the functions do not need to be linear).

A review of Partially Supervised Learning (PSL) approaches with special emphases on pattern classification and clustering is given in (Schwenker and Trentin, 2014). The major instances of PSL techniques are categorized into the following taxonomy: (1) active learning for training set design, where the learning algorithm has control over the training data; (2) learning from fuzzy labels, whenever multiple and discordant human experts are involved in the (complex) data labelling process; (3) SSL learning in pattern classification (further sorted out into: self-training, SSL with generative models, semi-supervised support vector machines; SSL with graphs); (4) SSL in data clustering, using additional constraints to incorporate expert knowledge into the clustering process; (5) PSL in ensembles and learning by disagreement; and (6) PSL in artificial neural networks. The paper also points out the main issues which are still open.

Note that some of the existing models cannot be applied to the present learning problem since they require all classifiers trained on the same dataset (or on examples from the same population, with the same dimensions and features). Other models do not take advantage of the order information in the classes to improve the generalization performance. Intuitively, they do not make the best usage of the information available in MOL for the learning process.

7.3 Learning max ordinal relations

Since a new learning concept is being explored, one option is to adapt existing types of models (e.g. ANNs, SVMs, etc.) to the new objective function. However, it would be interesting to accommodate this formulation under the ordinal class problem. This would allow the use of mature and optimized algorithms, developed for this well-established problem. Two alternative iterative processes are therefore discussed here that have, at the core, a base classifier for multi-class classification problems, which is not necessarily ordinal.

The proposed methodology makes use of a Base Classifier for each view. In general, the base classifiers can be from different types and they can be adapted for the data in the corresponding view (e.g., an SVM for the first view and a decision tree for the second). The methodologies to be presented do not make any assumptions in this respect. Regarding the ordinal nature of the data, the scenario is different. Both methodologies make use of the order in the splitting of the training dataset in two, according to the predictions on each view. Inside the framework, the Base Classifier may or may not take advantage of the order information. It is expected that Classifiers that do make use of the order nature of the classes (including 'more knowledge' in the learning process) achieve a better performance.

7.3.1 MOL.LA algorithm

Each base classifier is initialized by training it with all data from both the corresponding subset $\mathscr{S}_i$ and the subset $\mathscr{S}_{12}$. In the initialization, the labels in the subset $\mathscr{S}_{12}$ for each base classifier are

assumed to correspond to the observed labels. For all subsequent iterations, consider the objective function in Equation (7.1) rewritten as:

$$\{f^*, g^*\} = \arg\min_{f,g} \left(\sum_{\mathbf{x}_i \in \mathscr{S}_1} \mathscr{L}(f(\mathbf{x}_i), y_i) + \sum_{\mathbf{z}_i \in \mathscr{S}_2} \mathscr{L}(g(\mathbf{z}_i), y_i) + \sum_{\substack{\mathbf{x}_i, \mathbf{z}_i \in \mathscr{S}_{12} \\ f(\mathbf{x}_i) > g(\mathbf{z}_i)}} \mathscr{L}(f(\mathbf{x}_i), y_i) + \sum_{\substack{\mathbf{x}_i, \mathbf{z}_i \in \mathscr{S}_{12} \\ f(\mathbf{x}_i) < g(\mathbf{z}_i)}} \mathscr{L}(g(\mathbf{z}_i), y_i) \right) \tag{7.2}$$

where the last term in Equation (7.1) has been split in two, according to which of the views predicts the highest value.

Under a Local Approximation (LA) assumption that, in the next iteration, the order relation between the individual predictions $f(.)$ and $g(.)$ on the observations is kept, the current hypothesis is optimized (the terms classifier and hypothesis are here used interchangeably) by retraining the model f on the set of observations $\mathscr{S}_1 \cup \{\mathbf{x}_i, \mathbf{z}_i : \mathbf{x}_i, \mathbf{z}_i \in \mathscr{S}_{12} \wedge f(\mathbf{x}_i) > g(\mathbf{z}_i)\}$ and by retraining the model g on the set of observations $\mathscr{S}_2 \cup \{\mathbf{x}_i, \mathbf{z}_i : \mathbf{x}_i, \mathbf{z}_i \in \mathscr{S}_{12} \wedge f(\mathbf{x}_i) < g(\mathbf{z}_i)\}$. Although the points in $\mathscr{S}_{12}$ with $f(\mathbf{x}_i) = g(\mathbf{z}_i)$ are not explicitly addressed in the above description, they are randomly split between the datasets used to update the two models. Furthermore, note that with this approach, in each iteration, each point is used when updating one of the models but not in both.

As already mentioned, with MOL.LA any multi-class method without modifications can be selected for the base classifier. Naturally, it is expected that base classifiers adapted for ordinal data achieve better performance than conventional classifiers for nominal data. MOL.LA corresponds to the batch method initially proposed in (Cardoso and Domingues, 2011).

The adaptations of the MOL.LA algorithm for the two generalizations considered at the end of Section 7.1 are simple. In the extension of the model to M views, an observation can include any subset of views, corresponding to $2^M - 1$ different combinations. The base models are initialized as before, using the subset of observations containing the corresponding view and assuming that the label is due to that view. The iterative process also remains the same, where each model is re-trained with the subset of observations where the model prediction is maximal. In the extreme case where every observation includes all views, the process remains the same as before.

7.3.2 MOL.CD algorithm

An alternative approach is to consider a Coordinate Descent (CD) methodology. The base classifiers are initialized as before but now, in each iteration, two steps are performed:

- in the first step, the hypothesis $g(.)$ is fixed and the objective function is optimized over $f(.)$
- in the second step, the hypothesis $f(.)$ is fixed and the objective function is optimized over $g(.)$.

In the first step, since $g(.)$ is considered fixed, the optimization of Equation (7.1) is equivalent to:

$$f^* = \arg\min_f \left(\sum_{\mathbf{x}_i \in \mathscr{S}_1} \mathscr{L}(f(\mathbf{x}_i), y_i) + \sum_{\mathbf{x}_i, \mathbf{z}_i \in \mathscr{S}_{12}} \mathscr{L}(\max(f(\mathbf{x}_i), g(\mathbf{z}_i)), y_i) \right) \tag{7.3}$$

Splitting the last term in two gives,

$$f^* = \arg\min_f \left(\sum_{\mathbf{x}_i \in \mathscr{S}_1} \mathscr{L}(f(\mathbf{x}_i), y_i) + \sum_{\substack{\mathbf{x}_i, \mathbf{z}_i \in \mathscr{S}_{12} \\ f(\mathbf{x}_i) > g(\mathbf{z}_i)}} \mathscr{L}(f(\mathbf{x}_i), y_i) + \sum_{\substack{\mathbf{x}_i, \mathbf{z}_i \in \mathscr{S}_{12} \\ f(\mathbf{x}_i) < g(\mathbf{z}_i)}} \mathscr{L}(g(\mathbf{z}_i), y_i) \right) \tag{7.4}$$

Since y_i and $g(\mathbf{z}_i)$ are assumed known and fixed, predictions by the hypothesis f above $g(\mathbf{z}_i)$ are penalized according to the adopted loss function. Predictions by the hypothesis f below $g(\mathbf{z}_i)$ are penalized by $\mathscr{L}(g(\mathbf{z}_i), y_i)$, which is independent of $f(\mathbf{x}_i)$. Considering MAE as the loss function for illustration purposes, an error in an observation in $\mathscr{S}_1$ is penalized as depicted in Figure 7.2a. Note that, although $f(.)$ only assumes values in a finite set, a continuous representation was adopted in Figure 7.2 for better visualization. The loss in an observation in $\mathscr{S}_{12}$ depends on the relative values of y and $g(\mathbf{z})$, as represented in Figure 7.2b and Figure 7.2c.

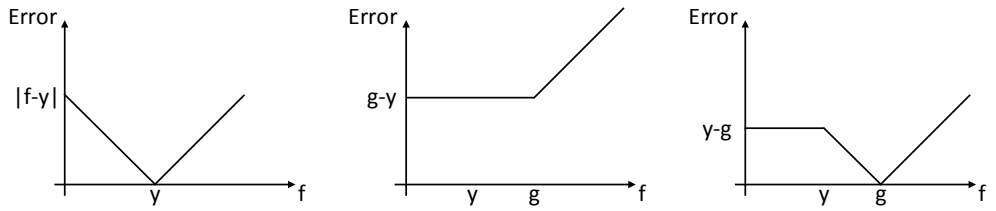


Figure 7.2: MAE functions when g is fixed: (a) subset $\mathscr{S}_1$, (b) subset $\mathscr{S}_{12}$ and $g > y$, (c) subset $\mathscr{S}_{12}$ and $g < y$ (similarly for the case when f is fixed).

It is important to emphasize that this learning process incorporates two notions of distance or error. First, the Base Classifier is internally optimizing some notion of error, typically defined in the continuous domain. For instance, when instantiated with SVMs, the internal loss is given by the hinge loss function. Externally, the framework is using a notion of loss defined in the space of the categorical values, $\mathscr{Y} = \{y_1, \cdots, y_K\}$, where $y_1 \prec \cdots \prec y_K$ and $\prec$ is a linear order relation in $\mathscr{Y}$. MAE is used as an example, but other options include using mean square error, Average MAE (AMAE) or the ordinal classification index (Cardoso and Sousa, 2011).

Note that with this approach, in each iteration, each model receives all points for training, although with different costs. The adaptations of the MOL.CD algorithm for the two generalizations considered at the end of section 7.1 are also simple. The extension to M views is accomplished by fixing all but one of the M models at a time. Therefore, in each iteration, a total of M steps is performed.

7.3.3 Summation

MOL.CD is a kind of coordinate descent in the space of the models. Like all gradient or coordinate descent methods, it can stay trapped in locally optimal solutions if the performance surface in the space of models is complex. In MOL.LA, the focus is on the partitioning of training instances into two subsets. Arguably, the most difficult part of the learning in MOL is understanding for which subset of points in $\mathscr{S}_{12}$ the label is due to the first view (subset $\mathscr{S}_{12}^{(1)}$) and for which subset of points in $\mathscr{S}_{12}$ the label is due to the second view (subset $\mathscr{S}_{12}^{(2)}$). When this is known, the learning problem is now equivalent to the training of two independent classifiers, one in $\mathscr{S}_1 \cup$ (subset $\mathscr{S}_{12}^{(1)}$), the other in $\mathscr{S}_1 \cup$ (subset $\mathscr{S}_{12}^{(2)}$) (or semi-supervised approaches, like TOR, which likely provide better solutions). With MOL.LA, the goal is to travel the space of partitions, choosing the next partition to be evaluated based on the predictions obtained by the classifiers trained in the current partition.

Both MOL.LA and MOL.CD are iterative methods that try to decrease the loss at each iteration. However, like many methods of this kind, the loss is not guaranteed to decrease monotonously. Moreover, for certain combinations of the loss function, base classifier and dataset the loss can fluctuate and the methods may not converge. Also, as is typical in these methods, they are either run for a pre-specified number of iterations, until there is no significant change in the models, or until the loss is below a pre-specified quality value.

The prediction stage is common to both frameworks. Note that the output of the learning process for both frameworks is a set of classifiers, one per view, able to make predictions when receiving as input the attributes of the corresponding view.

When in the presence of a test instance, possibly only a subset of the views is present. The predicted output for the test instance will be the maximum of the individual predictions for each of the views that are present in the instance.

7.4 Results

The data used in this experimental section is the one described in Chapter 6[1]. In short, $\mathscr{S}_1$ corresponds to 28 examples of mammograms with masses only, from which a set of 9 features per image was extracted when using the Ground Truth (GT) contours and 5 features per image when using CaPTOR. $\mathscr{S}_2$ corresponds to 234 examples of mammograms with calcifications only, from which a set of 7 features per image was extracted. The set $\mathscr{S}_{12}$ corresponds to a total of 79 images containing both masses and calcifications. In this case, a total of 16 (12 for CaPTOR) features were extracted per image, from which the first 9 (or 5) features try to characterize the mass and the remaining 7 characterize the calcifications in the same image. Histograms of the dataset representing the BI-RADS distribution can be seen in Figure 7.3.

[1] We refer to (Domingues and Cardoso, 2013b) for experiments on synthetic data.

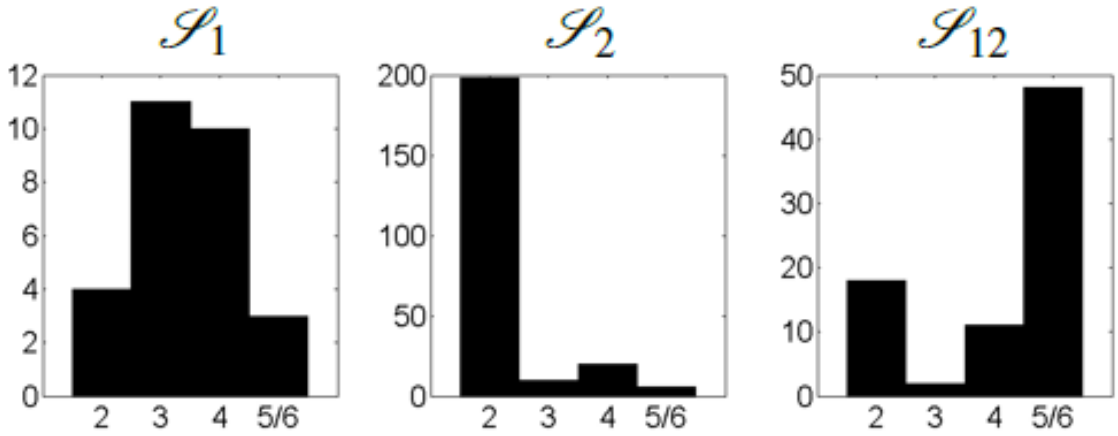


Figure 7.3: Dataset class distributions. Left: images with mass only; centre: images with calcifications only; Right: images with both mass and calcifications. X-axis represents the BI-RADS class.

The non-ordinal extension from binary to multi-class was conducted with the one-against-one method. The ordinal methods used were Frank&Hall and data replication[1]. All models were instantiated with SVMs. The MOL.CD framework was not instantiated in the nominal method, since it is not clear how to incorporate misclassification costs into this base model. With the data replication and the Frank&Hall methods, the misclassification cost of each observation is managed by controlling the presence of the observation in each data replica or in the training of each individual classifier.

For comparison purposes, both a Standard multi-class method (Cardoso and Domingues, 2011) and tri-training (Zhou and Li, 2005) were included in the experiments. In the Standard multi-class, for the two-view example, the subset $\mathscr{S}_{12}$ is ignored while two models are derived, one for each view. In the test phase, the maximum predicted by the two classifiers is taken as the final prediction. Standard multi-class models were tested, one-against-one SVMs, Frank&Hall, and data replication. Only the best results (namely data replication) are presented.

Table 7.1 summarizes the results for the baseline techniques, Table 7.2 contains the MOL.LA results, and Table 7.3 the MOL.CD. Ground Truth (GT) datasets correspond to dataset with features extracted from GT detections while for CaPTOR datasets mass features were extracted from contours generated by the CaPTOR algorithm. Subscripts refer to the feature selection technique used to rank and select the features (Chapter 6).

The statistical significance analysis of the complete test set results yielded:

- the use of mass contours retrieved by the CaPTOR algorithm in comparison with the use of GT segmentations showed no statistical difference for the set of features selected by *dcorr* or MIC. When using *corr* selected features, CaPTOR contours led to statistically better results

[1]TOR could also have been used. Its implementation is, however, non-trivial. In order not to bias the results due to implementation issues, the authors were contacted. However, the original implementation was not available for release. Although better results could have been achieved by instantiating MOL with TOR, it is considered that the results here presented are sufficient to prove the validity of the MOL framework. In the future, experiments will be made by instantiating MOL with a semi-supervised ordinal technique.

Table 7.1: MAE results for the BI-RADS classification problem: baseline techniques. Results are presented in the format "*complete test set* % ($\mathscr{S}_1$% / $\mathscr{S}_2$% / $\mathscr{S}_{12}$%)".

Dataset	Standard Model	Tri-Training
GT_{corr}	15 (23/3/118)	17 (28/4/134)
GT_{dcorr}	13 (15/2/116)	16 (29/2/139)
GT_{MIC}	14 (28/2/114)	16 (42/2/136)
$CaPTOR_{corr}$	13 (23/2/120)	16 (29/4/141)
$CaPTOR_{dcorr}$	14 (26/2/114)	15 (37/2/133)
$CaPTOR_{MIC}$	14 (18/3/112)	16 (29/3/136)

Table 7.2: MAE results for the BI-RADS classification problem: MOL.LA. Results are presented in the format "*complete test set* % ($\mathscr{S}_1$% / $\mathscr{S}_2$% / $\mathscr{S}_{12}$%)".

Dataset	Non-ordinal	Frank&Hall	Data Replication
GT_{corr}	10 (41/3/66)	9 (10/3/65)	7 (70/3/31)
GT_{dcorr}	9 (50/3/48)	9 (15/4/58)	10 (46/3/64)
GT_{MIC}	11 (50/5/55)	12 (20/7/58)	13 (49/7/63)
$CaPTOR_{corr}$	7 (49/2/48)	7 (14/2/59)	8 (47/3/48)
$CaPTOR_{dcorr}$	10 (45/5/49)	8 (16/2/52)	8 (46/3/47)
$CaPTOR_{MIC}$	14 (50/9/44)	12 (22/8/47)	13 (50/7/47)

Table 7.3: MAE results for the BI-RADS classification problem: MOL.CD. Results are presented in the format "*complete test set* % ($\mathscr{S}_1$% / $\mathscr{S}_2$% / $\mathscr{S}_{12}$%)".

Dataset	Frank&Hall	Data Replication
GT_{corr}	9 (10/3/65)	9 (40/3/58)
GT_{dcorr}	8 (15/3/58)	9 (47/2/63)
GT_{MIC}	8 (13/3/55)	10 (48/4/62)
$CaPTOR_{corr}$	7 (14/2/59)	7 (41/2/48)
$CaPTOR_{dcorr}$	7 (12/1/51)	8 (43/2/50)
$CaPTOR_{MIC}$	9 (20/4/46)	11 (49/5/49)

in the Standard Model, MOL.LA with non-ordinal and Frank&Hall instantiations and both MOL.CD instantiations [1]. For the other models, differences were not significant;

- no statistical difference was found between the use of *corr* or *dcorr* for feature selection, except for MOL.LA with Data replication when using the GT segmentation, where *corr* is significantly better than *dcorr*. *dcorr* was significantly better than MIC when using CaPTOR segmentations for the MOL.LA method instantiated both with Frank&Hall and Data replication and when using GT segmentations for the MOL.LA method instantiated with Data replication. In the other experiments there was no statistical difference between the use of *dcorr* or MIC. *corr* is significantly better than MIC when using GT segmentations for the MOL.LA method instantiated with Frank&Hall and Data replication and when using CaPTOR segmentations for the MOL.LA method instantiated with the non-ordinal technique, Frank&Hall and Data replication and for the MOL.CD method instantiated with Data replication;

- in all experiments, there was at least one instantiation with a base ordinal classifier of MOL.LA or MOL.CD that was statistically better than the two conventional methods;

[1] This is an unexpected result that will be further studied in future work.

- in five experiments (out of six), tri-training was statistically worse than all the other models. In the remaining experiment (dataset CaPTOR$_{MIC}$) the only non-significant difference was between tri-training and non-ordinal MOL.LA;
- in four experiments, the Standard Model was statistically worse than all the MOL.LA and MOL.CD instantiations. In all the six experiments, the Standard Model was statistically worse than all the MOL.CD instantiations;
- Frank&Hall and Data Replication in MOL.CD were statistically better than the corresponding MOL.LA instantiation in seven experiments (out of 12). MOL.CD was statistically worse than the corresponding MOL.LA instantiation in only one experiment. In the other experiments there was no statistical difference;
- there is no method which is statistically better than all the Frank&Hall and Data Replication instantiations, for any dataset.

Several aspects are worth noting. A first observation is that automatic segmentation does not seem to negatively affect classification results. The use of MIC for feature ranking is not advised in this particular application. Both *corr* and *dcorr* behave in a similar way, however since *corr* is the most simple and common technique, its use is recommended in a practical implementation.

Both the MOL.LA and MOL.CD techniques perform better than the baseline methods. A legitimate conclusion is that, in a specific application scenario, it is sufficient to test and compare MOL.LA and MOL.CD instantiated with ordinal methods since there is always an instantiation superior to the other models.

When comparing the two methodologies proposed in this work, MOL.LA and MOL.CD, it is observed that MOL.CD behaves better for some datasets, in particular for the datasets built with features selected by MIC. This performance advantage is counterbalanced by the increase in the time to design the models. Note that while with MOL.LA each observation is used to update one and only one of the models, with MOL.CD every observation is used by every model (albeit with different costs), which allows MOL.CD to make better use of the data.

Concerning the time performance, it is pertinent to first to point out that all the algorithms were implemented in Matlab and no attempt was made to optimize the running times. Having this in mind, the fastest algorithms are the baseline ones (Tri-training and Standard Model), followed by MOL.LA. MOL.CD was the slowest methodology. Moreover, the Data replication implementation is slower than the corresponding Frank&Hall. When using these methodologies in practice, the trade-off between accuracy and performance must be considered.

7.5 Conclusions

The typical learning settings already studied in the literature are not necessarily the most interesting for practical applications, since they may not represent well the information that is available. In this chapter the max-ordinal learning paradigm was presented, existing in between supervised

and semi-supervised classification. For every observation, some information about the label is available. However, in a subset of the examples, the knowledge is incomplete. This corresponds to the worst-case classification of the individual views of the example.

A formalization of the max-ordinal learning paradigm led to two new learning schemes. The experimental evaluation showed that the methodologies developed gave better results. The adaptation of the training to the max paradigm makes the classifiers more suitable for the MOL problem. In comparison, the "blind" use of traditional classifiers can obtain sub-optimal results. The experiments conducted underline the importance of including prior knowledge when designing a classifier.

In the practical application studied here, and taking as example one of the 40 repetitions with the dataset CaPTOR$_{corr}$, the Standard Model correctly classified 69 images with BI-RADS 2 and one image with BI-RADS 4; one image with BI-RADS 2 is classified as 3 and other two as BI-RADS 4; nine images with BI-RADS 5 are classified as 4 (MAE = 17%). MOL.LA Data Replication correctly classifies 71 images with BI-RADS 2 and nine images with BI-RADS 5, misclassifying as BI-RADS 5 two images, one whose real BI-RADS is 2 and other 4 (MAE = 5%).

Chapter 8

Integrated performance analysis

In the previous chapters, each processing step was evaluated by assuming that the preceding analysis was successful. Here, the impact of each block on the final result is assessed. As in the previous chapters, 40 splits of the dataset into train and test subsets were made. The remainder of this Chapter consists of the side-by-side evaluation of each block of the fully connected pipeline with its counterpart using the ground truth (GT) information.

8.1 Pectoral muscle segmentation

The SPPC method was chosen for pectoral muscle segmentation because it has been shown in Chapter 3 to behave better than SPLE when no user input is available. Results on some of the test images for one of the 40 splits are given in Figure 8.1. As the pectoral muscle segmentation does not depend on any preceding task, its performance is the same for GT and automatic columns of Table 8.1.

8.2 Screening

The Screening pipeline has two blocks, density classification and benign breast identification (accordingly to density). Density classification error changes from $0.1873\ (0.0636)$ to $0.1864\ (0.0638)$ when using SPPC pectoral muscle segmentation instead of the GT information. This difference is not statistically significant ($p = 0.9461$). Some examples of correctly and incorrectly classified images for one of the 40 splits are presented in Figure 8.2. The influence on the final screening algorithm of the use of the SPPC pectoral segmentation and the automatic density classification against the use of GT information is shown in Table 8.1. The differences in TP and FN are statistical significant ($p < 0.001$) while the differences for TN and FP are not ($p = 0.0695$). This means that when using the GT information, more images are correctly classified as suspicious than when using the fully automatic method. Consequently, the number of suspicious images incorrectly classified as normal is lower when using the GT information. The amount of normal images correctly classified is not affected by the use of the automatically extracted information. The number

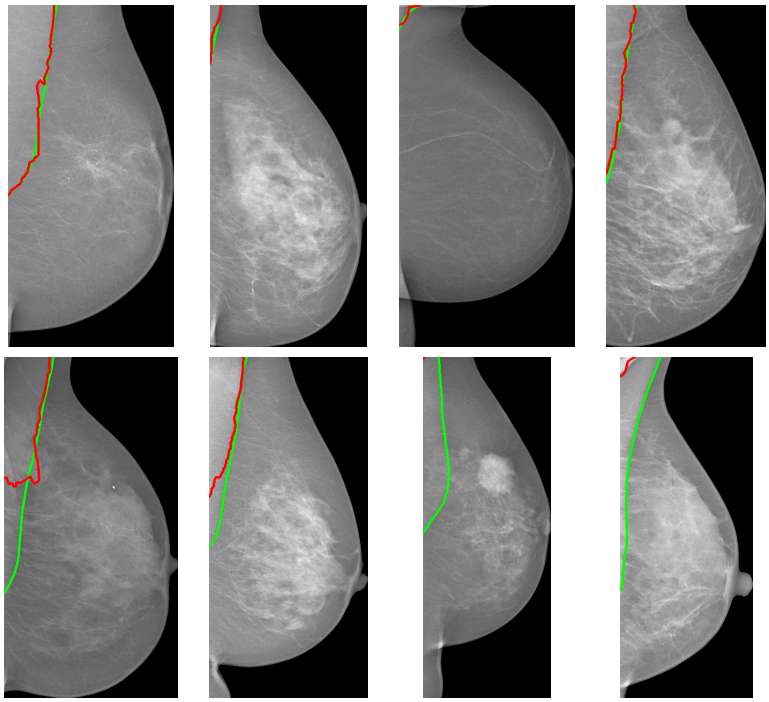


Figure 8.1: Pectoral muscle segmentation examples for one of the 40 splits. The GT is shown in green while automatic segmentation is plotted in red.

of normal images classified as suspicious is also not affected. Screening result examples are given in Figure 8.3.

8.3 Detection of suspicious regions

Calcification detection sensitivity increases from 56.4 (7.3) % to 63.8 (6.4) %, and the FPs increase from 47 (12) to 49 (12), as shown in Table 8.1. Both differences are statistically significant ($p < 0.0001$ and $p < 0.05$). In practice, this means that when using the fully connected pipeline, more detections are retrieved by the algorithm. While among these detections more calcifications are correctly located, some other spots that do not correspond to calcifications are also retrieved. Some illustrative examples are given in Figure 8.4.

A possible explanation for the fact that more calcifications are detected is that the automatic pectoral muscle segmentation tends to under-segment the muscle region (as shown in some of the examples of Figure 8.1) and consequently a greater proportion of the breast area is analysed when the pipeline is fully connected. In this way, more regions are considered as suspicious. In this way, real calcifications near the pectoral muscle border that are not detected when using the GT pectoral muscle segmentation, are now retrieved.

For the mass detection results, the differences in sensitivity are statistically significant ($p < 0.05$) while the differences in the number of FPs are not ($p = 0.8949$). Selected mass detection results are shown in Figure 8.5 where the individual Sensitivity and number of FPs per image are superimposed.

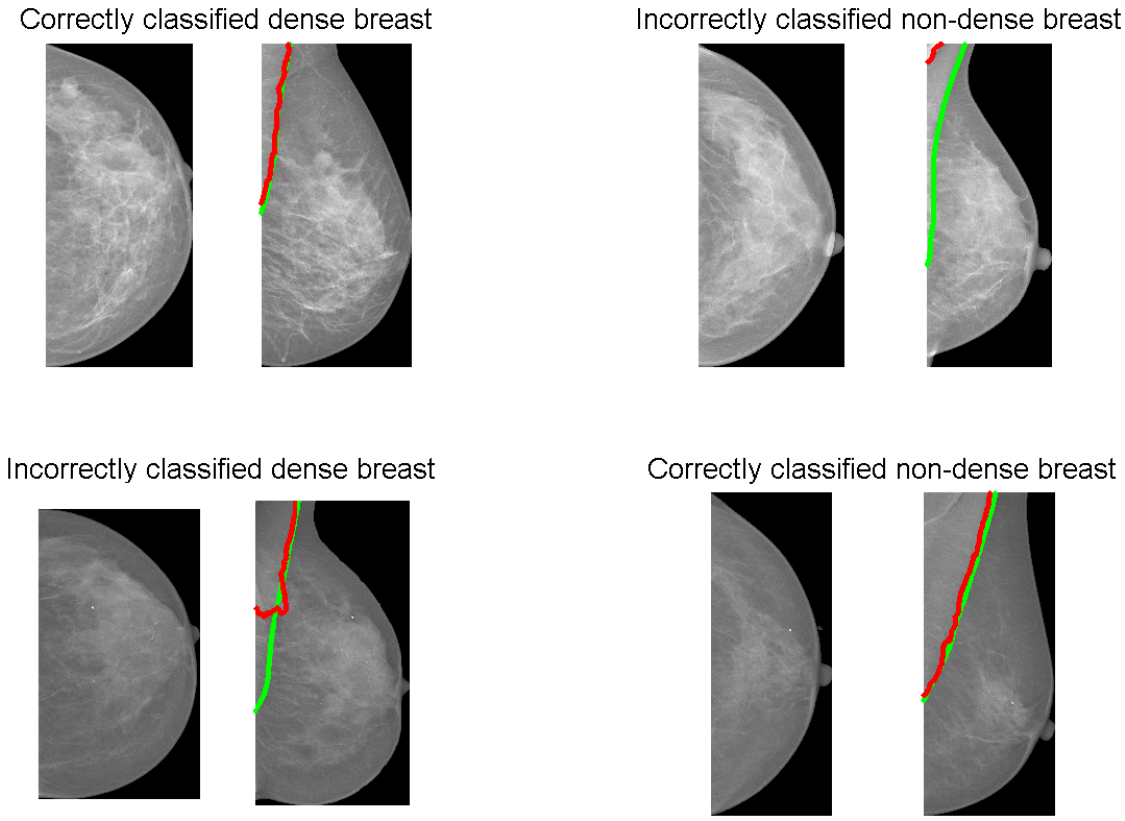


Figure 8.2: Examples of automatic density classification results for one of the 40 splits when using SPPC for pectoral muscle segmentation. The GT pectoral muscle contour is shown in green while SPPC result is presented in red.

Note that the statistical difference on sensitivity for calcifications ($p < 0.0001$) is stronger than the one for masses ($p < 0.05$). This is reinforced by the fact that the difference in FPs is significant for the calcifications case and not for the masses case. A possible explanation is that when doing automatic calcification detection there are two sources of errors, screening results and pectoral muscle segmentation. For the mass detection case, as explained in Section 5.5, the pectoral muscle information is not used and thus the only source of errors is the automatic screening.

8.4 BI-RADS classification

BI-RADS classification results presented in Table 8.1 were obtained using MOL.CD with Data replication and features selected by correlation. When using the automatic pipeline, the performance is significantly worse ($p < 0.0001$) than when using the GT information.

Normalized average values of the errors in the automatic classification are shown in Figure 8.6. As can be seen from the figure, the majority of BI-RADS classifications ($\approx 54\%$) do not change based on whether using the fully automatic approach compared to GT. Furthermore, an additional 20% only differ by one level.

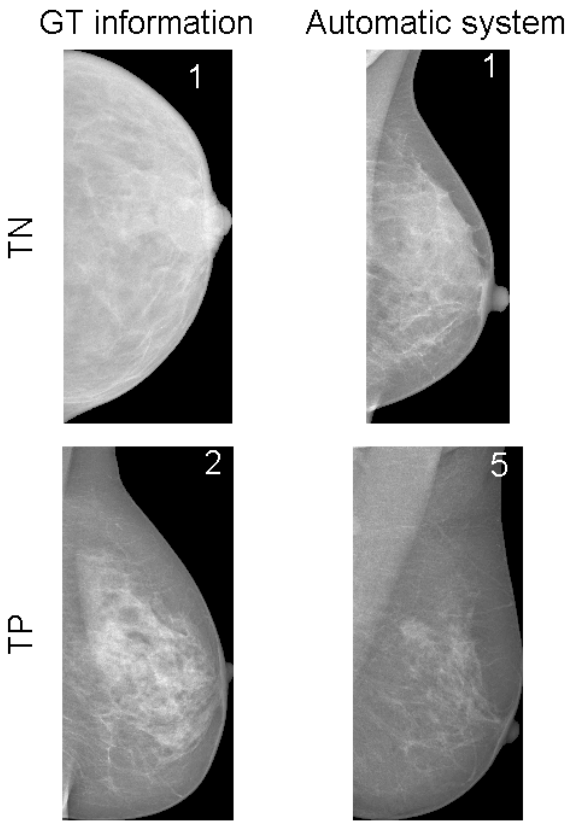


Figure 8.3: Examples of screening results for one of the 40 splits. The top row corresponds to TN cases while the bottom row refers to TP examples. In the left columns are shown examples of images correctly classified when using GT information but incorrectly classified by the complete automatic system. In the right column the only two examples that are correctly classified by the automatic system but incorrectly classified when using GT information are shown. BI-RADS GT information is superimposed in the respective images.

8.5 Overall results

Overall results are summarized in Table 8.1. Significantly better performances are shown in bold. It can be concluded that detection is the part of the pipeline that has a higher negative impact in the overall performance. All the other blocks do not significantly deteriorate the performance of the next block.

8.6 Conclusions

In this chapter the impact of the automatic algorithm on the overall performance was assessed. The main findings are:

- Screening
  - when using the GT information, more images are correctly classified as suspicious than when using the fully automatic method;

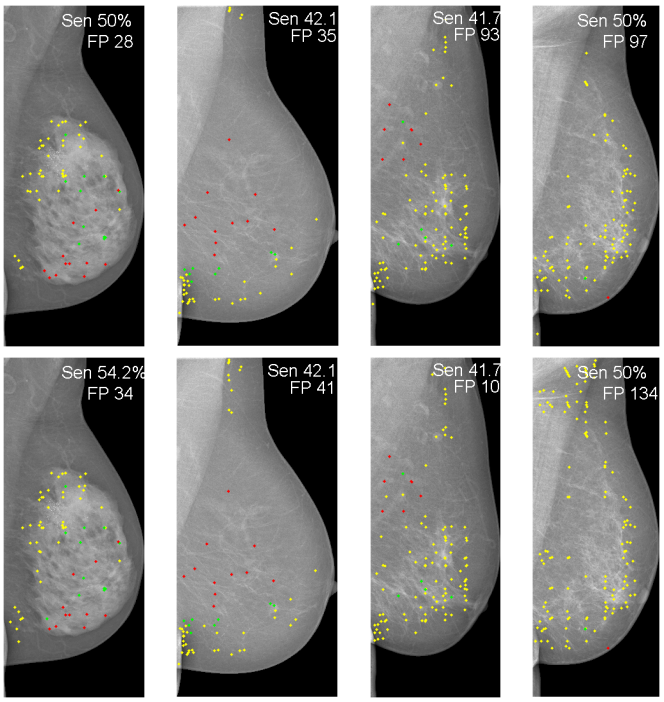


Figure 8.4: Calcification detection examples for one of the 40 splits. The top row corresponds to results using GT screening information, while the bottom row corresponds to results using the automatic pipeline. Green dots correspond to TPs, yellow to FPs and red to FNs.

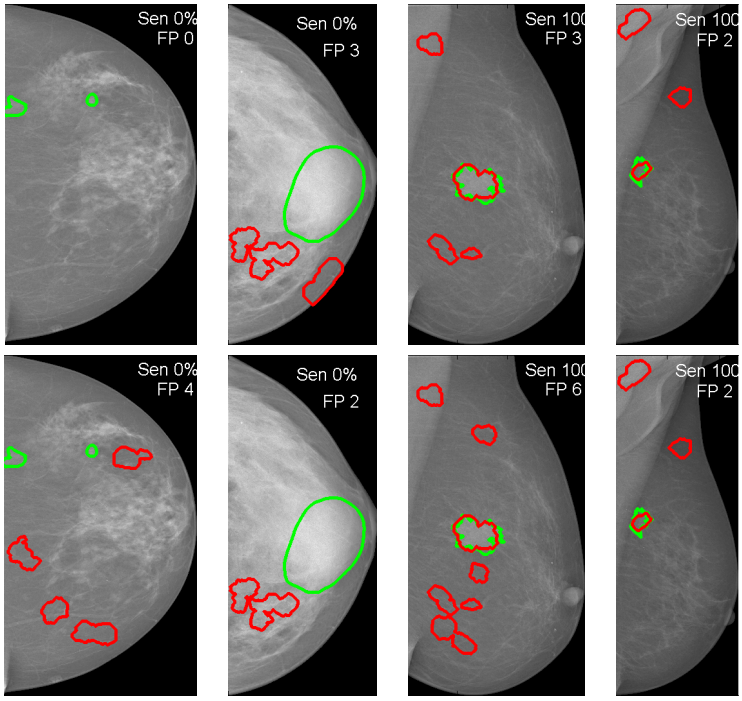


Figure 8.5: Mass detection examples for one of the 40 splits. The top row corresponds to results using GT screening information, while the bottom row corresponds to results using the automatic pipeline. The GT is shown in green while automatic detections are plotted in red.

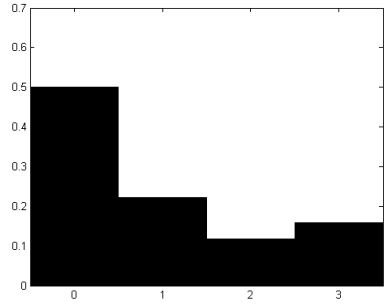


Figure 8.6: BI-RADS classification results. X-axis represents the absolute difference between the correct class and the predicted class. Y-axis corresponds to the normalized number of occurrences.

Table 8.1: Comparative analysis between the performance of each block using GT and the fully automatic system.

Component	GT	Automatic
Pectoral muscle (SPPC method)	AOM = 0.653 (0.285) CM = 0.767 (0.191) AD = 0.062 (0.082) AMED= 0.071 (0.088) HD = 0.173 (0.171)	
Screening (SVM RBF + Gabor features)	**TPr = 0.9175 (0.0435)** TNr= 0.1780 (0.1207) **FNr= 0.0825 (0.0435)** FPr= 0.8220 (0.1207)	TPr = 0.8246 (0.0244) TNr = 0.3272 (0.1758) FNr = 0.1754 (0.0244) FPr = 0.6728 (0.1758)
Calcification detection	sensitivity= 56.4 (7.3)% **FP= 47 (12)**	**sensitivity= 63.8 (6.4)%** FP= 49 (12)
Mass detection	sensitivity= 47.6 (9.2)% FP = 4 (0.3)	**sensitivity= 48.8 (9.1)%** FP= 4 (0.3)
BI-RADS classification	**MAE= 10%**	MAE= 88%

  - the number of suspicious images incorrectly classified as normal is lower when using the GT information;
  - the number of normal images correctly classified and the number of normal images classified as suspicious are not affected by the use of the automatically extracted information

- Detection
  - for calcifications, when using the fully connected pipeline, more detections are retrieved by the algorithm. In this way, more calcifications are correctly located, but also more spots that do not correspond to calcifications are retrieved
  - for the mass detection results, sensitivity also increases when using the automatic pipeline, while the number of FPs remains the same

- Classification
  - the automatic BI-RADS assessment suffers from the accumulated errors of the previous tasks

In an overall view it is believed by the author that pectoral muscle detection, screening, mass segmentation, feature extraction and BI-RADS classification are ready to be used in practice. The detection algorithms, however, need to be improved in order to provide higher sensitivities with fewer FPs.

Chapter 9

Conclusions

The present thesis described an effort to develop image and machine learning methods to help radiologists in the analysis of mammogram images. In this chapter a summary of the developed techniques and respective results is reported. Next, some comments on the practical use of the developed techniques are presented and finally some directions for future work are suggested.

9.1 Summary of contributions and results

Contributions were made in the different phases, including: (1) pre-processing, (2) screening, (3) detection of suspicious regions, (4) characterization of suspicious regions, and (5) classification. All the techniques were thoroughly evaluated using a database of full field digital mammogram images that, along with the images, contains meta-data information like breast density, BI-RADS assessment and very accurate segmentations of suspicious regions. This database is known as INbreast.

Image manipulation can have a strong impact on the success of subsequent tasks. Typical pre-processing applied to mammogram images is the removal of the pectoral muscle region. Two methods for segmentation of the pectoral muscle were presented in this thesis, namely polar coordinates and the shortest path (SPPC) and shortest path with endpoints learnt by SVMs (SPLE). Two region metrics: Area Overlap Measure (AOM) and a combined measure of under-segmentation, over-segmentation and AOM (CM); and three contour based metrics: Average Distance (AD), Average Minimum Euclidean Distance (AMED) and Hausdorff Distance (HD) were used to evaluate the results. SPPC yielded: AOM= 0.662, CM= 0.772, AD= 0.055, AMED= 0.063 and HD= 0.159. For SPLE the achieved results were: AOM= 0.597, CM= 0.722, AD= 0.056, AMED= 0.061 and HD= 0.153. Differences were shown not to be statistically significant.

After pectoral muscle delineation, the mammogram exam goes through the screening block. Since it has been observed that CAD performance depends on breast density, breasts are first classified as fatty or dense. Then, for each breast type, a specific classification block is designed to determine if the breast exam is suspicious. An extensive evaluation by testing a large set of features in combination with several classifiers was performed. The best density classification results

were achieved with a k-Nearest Neighbours (kNN) classifier and using a feature vector consisting of statistic features extracted from both views (MLO and CC). Suspiciousness classifiers were developed using kNN and Gabor features. A true positive (TP) rate of 95%, true negative (TN) rate of 11%, with false negative (FN) rate of 4% and false positive (FP) rate of 89% were achieved. These performances are comparable with those reported for human experts and the system could thus replace one of the specialists in the current double reading practice.

Based on the outcome of the screening stage, non-suspicious patients return to the normal screening program advised by different countries while suspicious exams are sent to diagnosis. During this analysis it may be useful to direct the attention of the specialist to regions of the image that may be problematic. The two most common findings seen in mammogram images are calcifications and masses. Due to their different characteristics (size, intensity, shape, border contrast, etc.) two different methods were used for each finding. An algorithm based on Bayesian surprise was developed for calcification detection, while an Iris filter followed by a closed contour segmentation method made in the original coordinate system was used to detect and segment masses. Calcification detection has a sensitivity of 51.9 (7.9) % with FP of 37 (11). Mass detection sensitivity decreases with the increase in breast density and increases with BI-RADS assessment. Detection is also significantly better for masses of sizes in the 10 to 20 mm range and worst for masses larger than 20 mm. Moreover, the algorithm is better at detecting non-spiculated masses than at detecting spiculated masses.

Calcifications and masses, when they exist, can be either benign or malign. BI-RADS describes important factors for malignancy determination including the distribution and morphology for calcifications and margins, shape and density for masses. A review of features used in the literature was presented. A lack of feature selection methods described in the literature for selection of continuous features for ordinal classification was observed. Hence, a large number of the existing features were evaluated on the INbreast database by using Pearson correlation ($corr$), distance correlation ($dcorr$) and the Maximal Information Coefficient (MIC). A subset of seven calcification features were shown to have a significant correlation with the BI-RADS final assessment. For masses, a final vector of nine features when using ground truth contours and five features when using CaPTOR was built.

The typical learning settings described in the literature are not necessarily the most interesting for practical applications, since they may not optimally represent the information that is available. For the specific classification problem studied in this thesis, a new learning paradigm is proposed, named max-ordinal learning (MOL), which sits in between supervised and semi-supervised classification. For every observation, some information about the label is available. However, in a subset of the examples, the knowledge is incomplete. This corresponds to the worst-case classification of the individual views of the example. A formalization of the max-ordinal learning paradigm led to two new learning schemes, MOL.LA and MOL.CD. MOL.CD uses coordinate descent in the space of the models, while in MOL.LA, the focus is on the partitioning of training instances into the two subsets. The experimental evaluation showed that the methodologies developed gave better results than traditional approaches. Mean Absolute Error (MAE) results were in the order

of 9% when using the subset of features selected by Pearson correlation.

All the described techniques were comprehensively evaluated both independently (assuming all the previous steps to be correct) and in connection. It could be concluded that pectoral muscle detection, screening, mass segmentation, feature extraction and BI-RADS classification are ready to be used in practice. The detection algorithms, however, need to be improved in order to provide higher sensitivities with fewer FPs.

The impact of the conducted research will be reflected in its ability to improve the quality of breast cancer detection, speeding up the time to output a diagnosis with the correspondent beneficial implications in treatment possibilities and psychological patient well being. The radiologist will also benefit from the fact that he can better use his time concentrating on more difficult cases.

9.2 Perspectives on practical applications

Concerning application scenarios it can be concluded that:

- Pectoral muscle segmentation performance is satisfactory and can be used in practice. Several options are available. It can be included in a software to help the MLO image acquisitions process. MLO view mammogram acquisition should follow several guidelines. One example is that the pectoral muscle should be imaged down to the mammilla level (Bulow et al., 2013). Pectoral muscle segmentation can be used together with a nipple detection technique to assess in real time if this (and other) guidelines are followed and, in case the image does not have the sufficient quality, the technician can be advised to reposition the patient and repeat the acquisition. Another option is to include the automatic segmentation in a CAD system. Knowledge about the pectoral muscle can be helpful in the image interpretation by a specialist. Finally, the algorithm can be used in a pipeline whose final goal is image registration, where typically some common landmarks are necessary in order to match two images;

- The automatic screening system can be used in the current double reading process by replacing one of the readers. Both the automatic system and a specialist would independently classify the images. In case the assessment is the same, the respective action is performed (either screening in a two year frame in case both assessments consider the exam as non-suspicious or the exam is sent to diagnosis in case both assessments consider the exam as suspicious). When the automatic result disagrees with the specialist opinion, the exam is forwarded to a panel of experts;

- Both mass and calcification detection techniques return a low sensitivity and a high number of FPs. They need thus to be improved in future work;

- Assuming a good detection (either manually or by improved automatic systems), feature extraction can be performed. In particular, in the case of masses, and for a manual detection, the user will only need to provide the mass centre and an approximate radius. The automatic

segmentation method can be used in practice both for visualization of the shape by the experts (e.g. included in a CAD system) and for shape feature extraction. Extracted features can be displayed to the user who, after some training, may be able to interpret the numeric features and make conclusions on calcification distribution and morphology, and the mass margins, shape and density characteristics, needed to make the final report in the BI-RADS standard;

- Automatic BI-RADS classification can also be included in a CAD system serving as a suggestion to the specialist who is going to perform the final report. Although the final assessment must always be made by the specialist, the automatic classification has shown an accuracy that can help the specialist when he is in doubt or can make the specialist interpret the exam more carefully in case the automatic class is higher than the one initially suggested by the specialist.

All of the above suggested scenarios have to be validated by clinical experts before implementation.

9.3 Future work

Each of the presented chapters can be further developed. On the pectoral muscle segmentation work, the step that seems to have the largest impact on the final segmentation result is the end-points detection. Therefore techniques for improved end-points detection are needed.

In the screening part, the results may be improved by extracting characteristics that better discriminate normal mammograms from suspicious images. The features extracted here are standard features that have been developed for other applications and were previously applied to this problem. There are, however, no specific features created for this particular application.

On the detection, and as also noted by other researchers, a higher sensitivity is achieved for calcifications than for masses. This comes at a cost that a high number of FPs are produced. This can reduce the trust of the specialist in the system. It is necessary to further develop the detection methods, making sure that no finding is missed but also limiting the number of FPs. Another important line of research that was not dealt with in this thesis is the architectural distortion detection. Architectural distortion is the third most common appearance of breast cancer. It is a distortion of the parenchymal architecture without a concomitant mass and it has been found to be the most challenging breast cancer manifestation to detect.

Concerning feature extraction, the creation of features specially designed for the Breast Imaging Reporting And Data System (BI-RADS) classification problem is envisioned. A different approach is to automatically learn the features from the images using, for instance, deep learning techniques. In order to pursue this second approach it is, however, necessary to have a very large repository of mammograms. The study and development of feature selection techniques specialized for the ordinal classification case is also anticipated.

One direction to continue the classification work is on the use of Reduction Techniques (RTs). Typically, RTs can be obtained by embedding the original problem in a higher-dimensional space. This embedding is implemented by replicating the training set points so that a copy of an original point is concatenated with extension feature vectors. The labels of the replicated points are set to maintain a particular structure in the extended space. This construction results in an artificially simpler problem, which is fed to a simpler learning algorithm. This idea has been applied to solve the Multiple Instance Learning (MIL) problem (Chen et al., 2006; Li and Yeung, 2009, 2010). By proceeding in a similar way, it may be possible to solve the Max-Ordinal Learning (MOL) problem using traditional ordinal learning methods. Moreover, RTs have been used to solve multi-class classification problems (El-Yaniv et al., 2008), including the ordinal setting (Cardoso and Costa, 2007) with a single binary classifier. Therefore, in the end, a single binary classifier could be used to address MOL.

Besides the improvement of the methods developed here, there are other possible research directions that have not been explored here. The use of ipsilateral (Kim et al., 2014b; Wu et al., 2014) and temporal (Kruger et al., 2010; Perez et al., 2014) information are two examples.

Another application of interest is to assist radiologist by presenting them with a set of images similar to the one being evaluated along with the known pathology of these reference images (Wang et al., 2014). This technique is known as Content-Based Image Retrieval (CBIR). While the potential of using similar cases to assist diagnosis has been demonstrated, a major challenge facing the CBIR approach is how to retrieve images with findings that are perceptually similar to the lesion under consideration (Wang et al., 2014). This is a possible future path of research.

To classify masses and calcifications according to their BI-RADS characteristics (calcification distribution, calcification morphology, calcification size, number of calcifications, mass location, mass size, mass morphology, mass margin and mass density) can also provide interesting information to display to the specialist.

Finally, although X-ray mammography is the most common cancer detection technique in several scenarios, including high-risk population, radiologists can also turn to multiple exams, like ultrasound and resonance imaging (MRI) to have a more reliable assessment. The development of automatic tools that integrate several breast image techniques can have an important impact in clinical practice but come at the cost of a new set of technical challenges, including image registration (Krüger et al., 2014; Mertzanidou et al., 2014).

Based on these many areas for future work, it is clear that there is much research still to be conducted to assist specialists in making an accurate assessment and early diagnosis of breast cancer.

References

ACR. *Mammography accreditation program*. American College of Radiology, 2014. URL `www.acr.org`.

P. Agrawal. *Contributions in computer assisted diagnosis: breast cancer and autoimmune diseases*. MSc dissertation, Indraprastha Institute of Information Technology New Delhi, 2014.

P. Agrawal, M. Vatsa, and R. Singh. Saliency based mass detection from screening mammograms. *Signal Processing*, 99:29–47, 2014.

F. Akram, J. H. Kim, and K. N. Choi. A preprocessing algorithm for the CAD system of mammograms using the active contour method. *Applied Medical Informatics*, 32(2), 2013.

S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 561–568, 2003.

H. R. Barde and S. D. Sawarkar. Cancer detection in mammogram using transform domain technique coupled with feature extraction. *International Journal of Computer Science and Management Research*, 2(4), 2013.

X. Bargallo, M. Velasco, G. Santamaria, M. Amo, P. Arguis, and S. S. Gomez. Role of computer-aided detection in very small screening detected invasive breast cancers. *Journal of Digital Imaging*, 26(3):572–577, 2013.

F. Baum, U. Fischer, S. Obenauer, and E. Grabbe. Computer-aided detection in direct digital full-field mammography: initial results. *European radiology*, 12(12):3015–3017, 2002.

S. Bessa, I Domingues, J. S. Cardoso, P. Passarinho, P. Cardoso, V. Rodrigues, and F. Lage. Normal breast identification in screening mammography: a study on 18 000 images. pages 325–330. International Conference on BioInformatics and BioMedicine (BIBM), 2014.

V. R. Borges, D. Guliato, C. A. Z. Barcelos, and M. A. Batista. An iterative fuzzy region competition algorithm for multiphase image segmentation. *Soft Computing*, pages 1–13, 2014.

J. Bozek, K. Delac, and M. Grgic. Computer-aided detection and diagnosis of breast abnormalities in digital mammography. In *50th IEEE International Symposium ELMAR*, volume 1, pages 45–52, 2008.

R. F. Brem and J. M. Schoonjans. Radiologist detection of microcalcifications with and without computer-aided detection: a comparative study. *Clinical radiology*, 56(2):150–154, 2001.

R. F. Brem, J. W. Hoffmeister, J. A. Rapelyea, G. Zisman, K. Mohtashemi, G. Jindal, M. DiSimio, and S. K. Rogers. Impact of breast density on computer-aided detection for breast cancer. *American Journal of Roentgenology (AJR)*, 184(2):439–444, 2005a.

R. F. Brem, J. W. Hoffmeister, G. Zisman, M. DeSimio, and S. K. Rogers. A computer-aided detection system for the evaluation of breast cancer by mammographic appearance and lesion size. *American Journal of Roentgenology (AJR)*, 184(3):893–896, 2005b.

R. F. Brem, J. A. Rapelyea, G. Zisman, J. W. Hoffmeister, and M. DeSimio. Evaluation of breast cancer with a computer-aided detection system by mammographic appearance and histopathology. *Cancer*, 104(5):931–935, 2005c.

T. Bulow, K. Meetz, D. Kutra, T. Netsch, R. Wiemker, M. Bergtholdt, J. Sabczynski, N. Wieberneit, M. Freund, and I. Schulze-Wenck. Automatic assessment of the quality of patient positioning in mammography. In *SPIE Medical Imaging*, volume 8670. International Society for Optics and Photonics, 2013.

K. Cadwell, C. Turner-Maffei, and B. O'Connor. *Maternal and infant assessment for breastfeeding and human lactation: a guide for the practitioner*. Jones and Bartlett Publishers, 2006. ISBN 9780763735777.

V. M. Campos, J. M. S. Martinez, C. B. Carron, J. A. Guirola, J. A. F. Gomez, and J. S. Perez. Tracks to face a breast imaging and succeed. In *European Society of Radiology (ECR)*, 2013.

Worldwide Breast Cancer. *Breast cancer statistics worldwide*. Worldwide Breast Cancer, 2009. URL `www.worldwidebreastcancer.com/learn/breast-cancer-statistics-worldwide`.

J. Canul-Reich and O. T. G. Mendez. A new collection of preprocessed digital mammograms. In *Advances in Artificial Intelligence and Its Applications*, pages 570–581. Springer, 2013.

J. S. Cardoso and J. F. P. Costa. Learning to classify ordinal data: the data replication method. *Journal of Machine Learning Research (JMLR)*, 8:1393–1429, 2007.

J. S. Cardoso and I. Domingues. Max-coupled learning: application to breast cancer. In *International Conference on Machine Learning and Applications (ICMLA)*, volume 1, pages 13–18, 2011.

J. S. Cardoso and R. Sousa. Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(8):1173–1195, 2011.

J. S. Cardoso, I. Domingues, I. Amaral, I. C. Moreira, P. Passarinho, J. S. Comba, R. Correia, and M. J. Cardoso. Pectoral muscle detection in mammograms based on polar coordinates and the shortest path. In *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4781–4784, 2010.

J. S. Cardoso, R. Sousa, and I. Domingues. Ordinal data classification using kernel discriminant analysis: a comparison of three approaches. In *International Conference on Machine Learning and Applications (ICMLA)*, pages 473–477, 2012.

J. S. Cardoso, I. Domingues, and H. P. Oliveira. Closed shortest path in the original coordinates with an application to breast cancer. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 2014.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

I. Cheikhrouhou, K. Djemal, D. Sellami, H. Maaref, and N. Derbel. New mass description in mammographies. In *First Workshops on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–5, 2008.

Y. Chen, J. Bi, and J. Z. Wang. MILES: multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(12):1931–1947, 2006.

Z. Chen, A. Oliver, E. Denton, and R. Zwiggelaar. Automated mammographic risk classification based on breast density estimation. *Pattern Recognition and Image Analysis*, 7887:237–244, 2013.

H. D. Cheng, X. Cai, X. Chen, L. Hu, and X. Lou. Computer-aided detection and classification of microcalcifications in mammograms: a survey. *Pattern Recognition*, 36(12):2967–2991, 2003.

M. P. Chhaya, F. P. W. Melchels, P. S. Wiggenhauser, J. T. Schantz, and D. W. Hutmacher. Breast reconstruction using biofabrication-based tissue engineering strategies. In *Biofabrication*, pages 183–216. William Andrew Publishing, 2013.

W. Chiracharit, Y. Sun, P. Kumhom, K. Chamnongthai, C. F. Babbs, and E. J. Delp. Normal mammogram detection based on local probability difference transforms and support vector machines. *Transactions on Information and Systems*, 1:258–270, 2007.

E. F. Conant and A. D. A. Maidment. Breast cancer imaging. *Science & Medicine*, 3(1):22–31, 1996.

I. Constantinou, M. Pattichis, C. Tziakouri, C. Pattichis, S. Petroudi, and C. Nicosia. Multiscale AM-FM models and instantaneous amplitude evaluation for mammographic density classification. Medical image understanding and analysis, 2014.

C. F. Cruz. *Automatic analysis of mammography images*. MSc dissertation, Faculdade de Engenharia da Universidade do Porto, 2011.

G. G. Davis. *Applied anatomy: the construction of the human body*. J. B. Lippincott Company, second edition, 1913. ISBN 9781409973881.

K. S. Deepak, N. V. K. Medathati, and J. Sivaswamy. Detection and discrimination of disease-related abnormalities based on learning normal cases. *Pattern Recognition*, 45(10):3707–3716, 2012.

E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1: 269–271, 1959.

I. Domingues. Multi-source automatic breast cancer diagnosis. Technical report, Faculdade Engenharia da Universidade do Porto, 2011.

I. Domingues and J. S. Cardoso. Max-coupled ordinal classification. In *17th Portuguese Conference on Pattern Recognition (RECPAD)*, 2011.

I. Domingues and J. S. Cardoso. Max ordinal learning applied to the BI-RADS classification of mammograms. In *1st PhD Students Conference in Electrical and Computer Engineering (StudECE)*, 2012.

I. Domingues and J. S. Cardoso. Mass detection on mammogram images: a first assessment of deep learning techniques. In *19th Portuguese Conference on Pattern Recognition (RECPAD)*, 2013a.

I. Domingues and J. S. Cardoso. Max ordinal learning. *IEEE Transactions on Neural Networks and Learning Systems*, 24, 2013b.

I. Domingues and J. S. Cardoso. Using bayesian surprise to detect calcifications in mammogram images. In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014.

I. Domingues, J. S. Cardoso, I. Amaral, I. C. Moreira, P. Passarinho, J. S. Comba, R. Correia, and M. J. Cardoso. Pectoral muscle detection in mammograms based on the shortest path with endpoints learnt by SVMs. In *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3158–3161, 2010.

I. Domingues, A. V. Alvarenga, J. S. Cardoso, and W. C. A. Pereira. SVM classification of breast tumours on ultrasound images using morphological features. In *XXIII, Congresso Brasileiro de Engenharia Biomedica (CBEB)*, 2012a.

I. Domingues, J. S. Cardoso, and P. Cardoso. Identification of benign breasts during mammogram screening. In *18th Portuguese Conference on Pattern Recognition (RECPAD)*, 2012b.

I. Domingues, E. Sales, J. S. Cardoso, and W. C. A. Pereira. INbreast database masses characterization. In *XXIII Congresso Brasileiro de Engenharia Biomedica (CBEB)*, 2012c.

A. R. Dominguez and A. K. Nandi. Detection of masses in mammograms using enhanced multilevel-thresholding segmentation and region selection based on rank. In *Fifth IASTED International Conference: biomedical engineering (BIEN)*, pages 370–375, 2007.

A. R. Dominguez and A. K. Nandi. Toward breast cancer diagnosis based on automated segmentation of masses in mammograms. *Pattern Recognition*, 42(6):1138–1148, 2009.

C. Dromain, C. Balleyguier, G. Adler, J. R. Garbay, and S. Delaloge. Contrast-enhanced digital mammography. *European Journal of Radiology*, 69(1):34–42, 2009. Breast Imaging.

F. Eddaoudi, F. Regragui, and K. Laraki. Characterization of the normal mammograms based on statistical features. In *International Symposium on Communications, Control and Signal Processing (ISCCSP)*, 2006.

R. El-Yaniv, D. Pechyony, and E. Yom-Tov. Better multi-class classification via a margin-optimized single binary problem. *Pattern Recognition Letters*, 29(14):1954–1959, 2008.

R. L. Ellis, A. A. Meade, M. A. Mathiason, K. M. Willison, and W. Logan-Young. Evaluation of computer-aided detection systems in the detection of small invasive breast carcinoma. *Radiology*, 245(1):88–94, 2007.

M. Elshinawy, A. Badawy, W. Abdelmageed, and M. Chouikha. Effect of breast density in selecting features for normal mammogram detection. In *International Symposium on Biomedical Imaging (ISBI)*, pages 141–147, 2011a.

M. Y. Elshinawy. *Pre-CAD normal mammogram detection algorithm based on tissue type*. PhD thesis, Howard University Washington, 2010.

M. Y. Elshinawy, W. W. Abdelmageed, A.-H. A. Badawy, and M. F. Chouikha. Pre-CAD system for normal mammogram detection using local binary pattern features. In *23rd IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 352–357, 2010a.

M. Y. Elshinawy, A.-H. A. Badawy, W. W. Abdelmageed, and M. F. Chouikha. Comparing one-class and two-class SVM classifiers for normal mammogram detection. In *39th IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7, 2010b.

M. Y. Elshinawy, A. H. A. Badawyy, W. W. Abdelmageed, and M. F. Chouikha. Normal mammogram detection using density information and texture features. In *SiiM*, 2011b.

M. Elter, C. Held, and T. Wittenberg. Contour tracing for segmentation of mammographic masses. *Physics in Medicine and Biology*, 55:5299, 2010.

D. R. Ericeira, A. C. Silva, A. C. Paiva, and M. Gattass. Detection of masses based on asymmetric regions of digital bilateral mammograms using spatial description with variogram and cross-variogram functions. *Computers in Biology and Medicine*, 43:987–999, 2013.

T. Esteves, P. Quelhas, A. M. Mendonca, and A. Campilho. Gradient convergence filters and a phase congruency approach for in vivo cell nuclei detection. *Machine Vision and Applications*, 23(4):623–638, 2012.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(9):1627–1645, 2010.

R. J. Ferrari, R. M. Rangayyan, J. E. L. Desautels, R. A. Borges, and A. F. Frere. Automatic identification of the pectoral muscle in mammograms. *IEEE Transactions on Medical Imaging (MedImg)*, 23(2):232–245, 2004.

C. I. Flowers, C. O'Donoghue, D. Moore, A. Goss, D. Kim, J.-H. Kim, S. G. Elias, J. Fridland, and L. J. Esserman. Reducing false-positive biopsies: a pilot study to reduce benign biopsy rates for BI-RADS 4A/B assessments through testing risk stratification and new thresholds for intervention. *Breast cancer research and treatment*, 139(3):769–777, 2013.

J. C. L. Fonseca. *Pre-CADs in breast cancer*. MSc dissertation, Faculdade de Engenharia da Universidade do Porto, 2013.

J. C. L. Fonseca, J. S. Cardoso, and I. Domingues. Pre-CADs in breast cancer. In *2nd PhD, Students Conference in Electrical and Computer Engineering (StudECE)*, 2013.

E. Frank and M. Hall. A simple approach to ordinal classification. In *12th European Conference on Machine Learning*, volume 1, pages 145–156. Springer, 2001.

G. Gallo and S. Pallottino. Shortest path algorithms. *Annals of Operations Research*, 13(1):1–79, 1988.

K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K. Ng. Computer-aided breast cancer detection using mammograms: a review. *IEEE Reviews in Biomedical Engineering*, 6:77–98, 2013a.

K. Ganesan, U. R. Acharya, K. C. Chua, L. C. Min, and K. T. Abraham. Pectoral muscle segmentation: a review. *Computer Methods and Programs in Biomedicine*, 110(1):48–57, 2013b.

K. Ganesan, U. R. Acharya, C. K. Chua, C. M. Lim, and K. T. Abraham. One-class classification of mammograms using trace transform functionals. *IEEE Transactions on Instrumentation and Measurement*, 63(2):304–311, 2014.

A. Garcia-Manso, C. J. Garcia-Orellana, H. M. Gonzalez-Velasco, R. Gallardo-Caballero, and M. Macias-Macias. Study of the effect of breast tissue density on detection of masses in mammograms. *Computational and Mathematical Methods in Medicine*, 2013.

D. Gaspar, L. Bre, N. Bischoff, and I. Domingues. Mammograms image processing. In *Investigacao Jovem na U.Porto, (IJUP)*, 2011.

C. H. Gils, J. D. M. Otten, A. L. M. Verbeek, and J. H. C. L. Hendriks. Mammographic breast density and risk of breast cancer: masking bias or causality? *European Journal of Epidemiology*, 14(4):315–320, 1998.

I. Gkioulekas, G. Evangelopoulos, and P. Maragos. Spatial bayesian surprise for image saliency and quality assessment. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 1081–1084, 2010.

O. H. Graven and D. A. H. Samuelsen. Tutorial project for a joint module in smart systems for bachelor students in electrical and computing engineering. In *IEEE Global Engineering Education Conference (EDUCON)*, pages 1–6, 2012.

V. L. Green and P. M. Weiss. *Breast disorders, an issue of obstetric and gynecology clinics*. Obstetrics and Gynecology Clinics of North America. Elsevier - Health Sciences Division, 2013. ISBN 9780323188647.

V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

R. M. Haralick, K. Shanmugam, and I. H. Dinstein. Textural feature for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3:610–621, 1973.

M. Heath, K. W. Bowyer, D. Kopans, W. P. Kegelmeyer, R. Moore, K. Chang, and S. MunishKumaran. Current status of the digital database for screening mammography. In *Digital Mammography*, pages 457–460, 1998.

B. Hela, M. Hela, H. Kamel, B. Sana, and M. Najla. Breast cancer detection: a review on mammograms analysis techniques. In *10th IEEE International Multi-Conference on Systems, Signals & Devices (SSD)*, pages 1–6, 2013.

M. A. Helvie, J. T. Chang, R. E. Hendrick, and M. Banerjee. Reduction in late-stage breast cancer incidence in the mammography era: implications for overdiagnosis of invasive cancer. *Cancer*, 2014.

S. H. Heywang-Koebrunner and I. Schreer. *Diagnostic breast imaging: mammography, sonography, magnetic resonance imaging and interventional procedures*. Thieme, 2014. ISBN 3131504110.

W. H. Hindle. *Breast care: a clinical guidebook for women's primary health care providers*. Springer, 1999. ISBN 9780387983486.

B. K. B. Hirata, J. M. M. Oda, R. L. Guembarovski, C. B. Ariza, C. E. C. Oliveira, and M. A. E. Watanabe. Molecular markers for breast cancer: prediction on tumour behaviour. *Disease markers*, 2014.

C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, National Taiwan University, 2003.

M. K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962.

Y.-J. Huang, D.-Y. Chan, D.-C. Cheng, Y.-J. Ho, P.-P. Tsai, W.-C. Shen, and R.-F. Chen. Automated feature set selection and its application to MCC identification in digital mammograms for breast cancer detection. *Sensors*, 13(4):4855–4875, 2013.

R. Hupse, M. Samulski, M. B. Lobbes, R. M. Mann, R. Mus, G. J. Heeten, D. Beijerinck, R. M. Pijnappel, C. Boetes, and N. Karssemeijer. Computer-aided detection of masses at mammography: interactive decision support versus prompts. *Radiology*, 266(1):123–129, 2013.

M. Hussain, S. Khan, G. Muhammad, I. Ahmad, and G. Bebis. Effective extraction of gabor features for false positive reduction and mass classification in mammography. *Applied Mathematics & Information Sciences*, 412(1):397–412, 2014.

D. M. Ikeda, R. L. Birdwell, K. O'Shaughnessy, E. A. Sickles, and R. J. Brenner. Computer-aided detection output on 172 subtle findings on normal mammograms previously obtained in women with breast cancer detected at follow-up screening mammography. *Radiology*, 230(3):811–819, 2004.

L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems (NIPS)*, pages 547–554, 2005.

A. Jalalian, S. B. T. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical Imaging*, 37(3):420–426, 2013.

H. Jing, Y. Yang, and R. M. Nishikawa. Retrieval boosted computer-aided diagnosis of clustered microcalcifications for breast cancer. *Medical Physics*, 39(2):676–685, 2012.

C. Juarez-Landin, V. Ponomaryov, J. L. Sanchez-Ratnirez, M. Martinez-Reyes, and V. Kravchenko. Wavelets based on atomic function used in detection and classification of masses in mammography. In A. Gelbukh and E. F. Morales, editors, *Advances in artificial intelligence (MICAI)*, volume 5317, pages 295–304, 2008.

N. Karssemeijer and G. T. Brake. Combining single view features and asymmetry for detection of mass lesions. In *Digital Mammography*, pages 95–102. Springer, 1998.

N. Karssemeijer and J. H. C. L. Hendriks. Computer-assisted reading of mammograms. *European radiology*, 7(5):743–748, 1997.

M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

K. M. Kelly, J. Dean, W. S. Comulada, and S.-J. Lee. Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts. *European radiology*, 20(3):734–742, 2010.

J. H. Kim, J. H. Cha, N. Kim, Y. Chang, M.-S. Ko, Y.-W. Choi, and H. H. Kim. Computer-aided detection system for masses in automated whole breast ultrasonography: development and evaluation of the effectiveness. *Ultrasonography*, pages 1–11, 2014a.

J. H. Kim, J. K. Ryu, G.-H. Jahng, and J. Y. Song. Double inversion recovery MR imaging of the breast: efficacy in detection of breast cancer. *Journal of magnetic resonance imaging (JMRI)*, 39(1):51–58, 2014b.

H. Kobatake. A convergence index filter for vector fields and its application to medical image processing. *Electronics and communications in Japan (Part III: fundamental electronic science)*, 89(6):34–46, 2006.

H. Kobatake, H. Takeo, and S. Nawano. Tumour detection system for full-digital mammography. In *Digital Mammography*, pages 87–94. Springer, 1998.

P. Kovesi. Image features from phase congruency. *Journal of Computer Vision Research*, 1(3), 1999.

E. Kozegar, M. Soryani, B. Minaei, and I. Domingues. Assessment of a novel mass detection algorithm in mammograms. *Journal of cancer research and therapeutics*, 9(4):592–600, 2013.

J. Krüger, J. Ehrhardt, A. Bischof, and H. Handels. Simulation of mammographic breast compression in 3D MR images using ICP-based b-spline deformation for multimodality breast cancer diagnosis. *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, 2014.

R. A. Kruger, R. B. Lam, D. R. Reinecke, S. P. Rio, and R. P. Doyle. Photoacoustic angiography of the breast. *Medical Physics*, 37(11):6096–6100, 2010.

C. Kuhl and M. C. Mahoney. *Breast MRI: expert consult*. Elsevier Health Sciences, 2013. ISBN 0323186467.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1–2):83–97, 1955.

M. Kuse, Y.-F. Wang, V. Kalasannavar, M. Khan, and N. Rajpoot. Local isotropic phase symmetry measure for detection of beta cells and lymphocytes. *Journal of Pathology Informatics*, 2, 2011.

S. M. Kwok, R. A. Chandrasekhar, Y. Attikiouzel, and M. T. Rickard. Automatic pectoral muscle segmentation on mediolateral oblique view mammograms. *IEEE Transactions on Medical Imaging (MedImg)*, 23(9):1129–1140, 2004.

Y. Lan, H. Ren, and J. Wan. A hybrid classifier for mammography CAD. In *Fourth IEEE International Conference on Computational and Information Sciences (ICCIS)*, pages 309–312, 2012.

R. Lavanya, N. Nagarajan, and M. N. Devi. False positive reduction in computer aided detection of mammographic masses using canonical correlation analysis. *Journal of Theoretical and Applied Information Technology*, 59(1):139–145, 2014.

C. H. Lee, D. D. Dershaw, D. Kopans, P. Evans, B. Monsees, D. Monticciolo, R. J. Brenner, L. Bassett, W. Berg, S. Feig, E. Hendrick, E. Mendelson, C. Orsi, E. Sickles, and L. W. Burhenne. Breast cancer screening with imaging: recommendations from the society of breast imaging and the ACR on the use of mammography, breast MRI, breast ultrasound and other technologies for the detection of clinically occult breast cancer. *Journal of the American College of Radiology*, 7(1):18–27, 2010.

W.-J. Li and D. Y. Yeung. Localized content-based image retrieval through evidence region identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1666–1673, 2009.

W.-J. Li and D. Y. Yeung. MILD: multiple-instance learning via disambiguation. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(1):76–89, 2010.

Y. Li, H. Chen, Y. Yang, and N. Yang. Pectoral muscle segmentation in mammograms based on homogeneous texture and intensity deviation. *Pattern Recognition*, 46(3):681–691, 2013.

L. Liu, Q. Liu, and W. Lu. Pectoral muscle detection in mammograms using local statistical features. *Journal of digital imaging*, 2014.

S. Liu. *The analysis of digital mammograms: spiculated tumour detection and normal mammogram characterization*. PhD thesis, Purdue University, 1999.

R. Llobet, M. Pollan, J. Anton, J. Miranda-Garcia, M. Casals, I. Martinez, F. Ruiz-Perales, B. Perez-Gomez, D. Salas-Trejo, and J.-C. Perez-Cortes. Semi-automated and fully automated mammographic density measurement and breast cancer risk prediction. *Computer methods and programs in biomedicine*, pages 1–11, 2014.

S.-C. B. Lo, J.-S. J. Lin, M. T. Freedman, and S. K. Mun. Application of artificial neural networks to medical image pattern recognition: detection of clustered microcalcifications on mammograms and lung cancer on chest radiographs. *Journal of VLSI, signal processing systems for signal, image and video technology*, 18(3):263–274, 1998.

M. Lobbes, M. Smidt, K. Keymeulen, R. Girometti, C. Zuiani, R. Beets-Tan, J. Wildberger, and C. Boetes. Malignant lesions on mammography: accuracy of two different computer-aided detection systems. *Clinical Imaging*, 37(2):283–288, 2013.

M. A. G. Lopez, N. G. Posada, D. C. Moura, R. R. Pollan, J. M. F. Valiente, C. S. Ortega, M. R. Solar, G. Diaz-Herrero, I. M. A. P. Ramos, J. P. Loureiro, T. C. Fernandes, and B. M. F. Araujo. BCDR: a breast cancer digital repository. In *15th International Conference on Experimental Mechanics (ICEM15)*, 2012.

LPCC. *Programa de rastreio de cancro da mama da Liga Portuguesa Contra o Cancro*. Liga Portuguesa Contra o Cancro (LPCC), 2009. URL `www.ligacontracancro.pt`.

A. Malich, C. Marx, M. Facius, T. Boehm, M. Fleck, and W. A. Kaiser. Tumour detection rate of a new commercially available computer-aided detection system. *European radiology*, 11(12): 2454–2459, 2001.

A. Malich, D. Sauner, C. Marx, M. Facius, T. Boehm, S. O. Pfleiderer, M. Fleck, and W. A. Kaiser. Influence of breast lesion size and histologic findings on tumour detection rate of a computer-aided detection system. *Radiology*, 228(3):851–856, 2003.

G. L. Masala. Computer aided detection on mammography. *World Academy of Science, Engineering and Technology*, 15(1):1–6, 2006.

B. Matheus, J. Neto, and H. Schiabel. Clustered microcalcification detection scheme for mammographic images. In *17th International Conference On Image Processing, Computer Vision, & Pattern Recognition (IPCV)*, pages 904–90, 2013a.

B. Matheus, H. Schiabel, and S. Carlos. A CADx scheme in mammography: considerations on a novel approach. In *Seventh International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP)*, pages 15–18, 2013b.

B. Matheus, L. B. Vercosa, B. Barufaldi, and H. Schiabel. Study of quality perception in medical images based on comparison of contrast enhancement techniques in mammographic images. In *SPIE Medical Imaging*, volume 9037. International Society for Optics and Photonics, 2014.

A. Mencattini, G. Rabottino, M. Salmeri, and R. Lojacono. An iris detector for tumoral masses identification in mammograms. In *IEEE International Workshop on Medical Measurements and Applications (MeMeA)*, pages 215–218, 2009.

T. Mertzanidou, J. Hipwell, S. Johnsen, L. Han, B. Eiben, Z. Taylor, S. Ourselin, H. Huisman, R. Mann, U. Bick, N. Karssemeijer, and D. Hawkes. MRI to x-ray mammography intensity-based registration with simultaneous optimisation of pose and biomechanical transformation parameters. *Medical image analysis*, 18(4):674–683, 2014.

M. G. Mini. Neural network based classification of digitized mammograms. In *Second Kuwait Conference on e-Services and e-Systems (KCESS)*, 2011.

M. G. Mini and T. Thomas. A neural network method for mammogram analysis based on statistical features. In *Conference on Convergent Technologies for the Asia-Pacific Region (TENCON)*, volume 4, pages 1489–1492, 2003.

K. Moberg, N. Bjurstam, B. Wilczek, L. Rostgaard, E. Egge, and C. Muren. Computed assisted detection of interval breast cancers. *European Journal of Radiology*, 39(2):104–110, 2001.

J. P. S. F. Monteiro. *Computer Aided Detection in Mammography*. MSc dissertation, Faculdade de Engenharia da Universidade do Porto, 2011.

I. C. Moreira. *Development and assessment of an e-learning course on senology for radiographers: a stratified randomized control trial*. MSc dissertation, FMUP&FCUP, 2012.

I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso. INbreast: towards a full field digital mammographic database. *Academic radiology*, 19:236–248, 2012.

E. Morris and L. Liberman. *Breast MRI: diagnosis and intervention*. Developments in cardiovascular medicine. Springer, 2005. ISBN 9780387219974.

D. C. Moura and M. A. G. Lopez. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, 8(4):561–574, 2013.

D. C. Moura, M. A. G. Lopez, P. Cunha, N. G. Posada, R. R. Pollan, I. Ramos, J. P. Loureiro, I. C. Moreira, B. M. F. Araujo, and T. C. Fernandes. Benchmarking datasets for breast cancer computer-aided diagnosis (CADx). In *Progress in Pattern Recognition, Image Analysis, Computer Vision and Applications*, pages 326–333. Springer, 2013.

R. Murakami, S. Kumita, H. Tani, T. Yoshida, K. Sugizaki, T. Kuwako, T. Kiriyama, K. Hakozaki, E. Okazaki, K. Yanagihara, S. Iida, S. Haga, and S. Tsuchiya. Detection of breast cancer with a computer-aided detection applied to full-field digital mammography. *Journal of Digital Imaging*, 26(4):768–773, 2013.

I. E. Naqa and Y. Yang. Techniques in the detection of microcalcification clusters in digital mammograms. In *Medical imaging systems technology: methods in diagnosis optimization*, pages 15–36. World Scientific, 2005.

T. Netsch and H.-O. Peitgen. Scale-space signatures for the detection of clustered microcalcifications in digital mammograms. In *IEEE Transactions on Medical Imaging*, volume 18, pages 774–786, 1999.

M. P. Nguyen, Q. D. Truong, D. T. Nguyen, T. D. Nguyen, and V. D. Nguyen. An alternative approach to reduce massive false positives in mammograms using block variance of local coefficients features and support vector machine. *Procedia computer science*, 20:399–405, 2013.

NHSBSP. Computer-aided detection in mammography. Quality Assurance Coordinating Group NHSBSP publication no. 48, National Health Service Breast Screening Programme (NHSBSP), 2001.

NIH. Breast cancer basics and you. *Medline Plus*, pages 20–23, 2014.

R. Nithya and B. Santhi. Mammogram analysis based on pixel intensity mean features. *Journal of computer science*, 8(3):329–332, 2012.

H. P. Oliveira. *An affordable and practical 3D solution for the aesthetic evaluation of breast cancer conservative treatment*. PhD thesis, Universidade do Porto, Faculdade de Engenharia, 2013.

H. P. Oliveira, J. S. Cardoso, A. Magalhaes, and M. J. Cardoso. Methods for the aesthetic evaluation of breast cancer conservation treatment: a technological review. *Current Medical Imaging Reviews*, 9(1):32–46, 2013.

A. Oliver. *Automatic mass segmentation in mammographic images*. PhD thesis, Universitat de Girona (Espana), 2008.

A. Oliver, J. Freixenet, R. Marti, and R. Zwiggelaar. A comparison of breast tissue classification techniques. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 4191, pages 872–879, 2006.

A. Oliver, J. Freixenet, J. Marti, E. Perez, J. Pont, E. R. E. Denton, and R. Zwiggelaar. A review of automatic mass detection and segmentation in mammographic images. *Medical Image Analysis*, 14:87–110, 2010.

C. S. Ortega, J. M. F. Valiente, M. R. Solar, G. D. Herrero, R. R. Pollan, M. A. G. Lopez, N. G. Posada, D. C. Moura, P. Cunha, I. Ramos, and J. Loureiro. Improving the breast cancer diagnosis using digital repositories. In *International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO)*, pages 571–578, 2013.

N. Otsu. A threshold selection method from grey-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.

D. C. Pereira, R. P. Ramos, and M. Z. Nascimento. Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm. *Computer Methods and Programs in Biomedicine*, 114(1):88–101, 2014.

S. Perez, M. J. Galon, Y. Andreu, E. Ibanez, E. Dura, A. Conchado, and E. Cardena. Posttraumatic stress symptoms in breast cancer patients: temporal evolution, predictors and mediation. *Journal of Traumatic Stress*, 27(2):224–231, 2014.

S. Petroudi and M. Brady. Breast density dependent computer aided detection. In *8th international conference on Digital Mammography (IWDM)*, volume 4046, pages 34–38, 2006.

D. L. Pham, C. Xu, and J. L. Prince. Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2(1):315–337, 2000.

T. Podsiadly-Marczykowska, A. Wroblewska, and A. Przelaskowski. Classification of microcalcifications into BI-RADS morphologic categories: preliminary results. *Biocybernetics and Biomedical Engineering*, 29(4):83–93, 2009.

K. Polyak. Breast cancer: origins and evolution. *Journal of Clinical Investigation*, 117(11):3155–3163, 2007.

J. M. S. Prewitt. Object enhancement and extraction. *Picture processing and Psychopictorics*, 10 (1):15–19, 1970.

B. Qi, H. Li, X. Pan, and Y. Kang. Breast mass segmentation in mammography using improved dynamic programming. In *IEEE International Conference on Information and Automation (ICIA)*, pages 215–220, 2012.

V. Raman, P. Sumari, H. H. Then, and S. A. K. Al-Omari. Review on mammogram mass detection by machine learning techniques. *International Journal of Computer and Electrical Engineering*, 3(6), 2011.

R. M. Rangayyan, N. Mudigonda, and J. Desautels. Boundary modelling and shape analysis methods for classification of mammographic masses. *Medical and Biological Engineering and Computing*, 38(5):487–496, 2000.

R. M. Rangayyan, F. J. Ayres, and J. E. L. Desautels. A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs. *Journal of the Franklin Institute*, 344(3–4): 312–348, 2007.

M. A. Rawashdeh, R. M. Bourne, E. A. Ryan, W. B. Lee, M. W. Pietrzyk, W. M. Reed, N. Borecky, and Patrick C. Brennan. Quantitative measures confirm the inverse relationship between lesion spiculation and detection of breast masses. *Academic radiology*, 20(5):576–580, 2013.

D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

P. B. Ribeiro. *Esquema CADx para Classificacao Nodulos em Imagens Mamograficas Digitais Baseado na Segmentacao pelo Modelo EICAMM*. PhD thesis, Universidade Sao Paulo, USP, Brasil, 2013.

M. Rizzi, D. A. Matteo, and B. Castagnolo. Review: health care CAD systems for breast microcalcification cluster detection. *Journal of Medical and Biological Engineering*, 32(3), 2012.

A. Rojas-Dominguez and A. K. Nandi. Development of tolerant features for characterization of masses in mammograms. *Computers in Biology and Medicine*, 39(8):678–688, 2009.

L. C. S. Romualdo, M. A. C. Vieira, H. Schiabel, N. D. A. Mascarenhas, and L. R. Borges. Mammographic image de-noising and enhancement using the anscombe transformation, adaptive wiener filtering and the modulation transfer function. *Journal of Digital Imaging*, 26(2): 183–197, 2013.

C. J. Rose. *Statistical models of mammographic texture and appearance*. PhD thesis, University of Manchester, 2005.

P. L. Rosin. Classification of pathological shapes using convexity measures. *Pattern Recognition Letters*, 30(5):570–578, 2009.

Y.-X. Ruan, H.-T. Lin, and M.-F. Tsai. Improving ranking performance with cost-sensitive ordinal classification via regression. *Information Retrieval*, 17(1):1–20, 2014.

M. P. Sampat, A. C. Bovik, and M. K. Markey. Classification of mammographic lesions into BI-RADS shape categories using the beamlet transform. In J. M. Fitzpatrick and J. M. Reinhardt, editors, *SPIE*, volume 5747, pages 16–25, 2005a.

M. P. Sampat, M. K. Markey, and A. C. Bovik. Computer-aided detection and diagnosis in mammography. In *Handbook of image and video processing*, volume 2.1, pages 1195–1217. Academic Press, 2005b.

M. P. Sampat, P. Mehul, A. C. Bovik, C. Alan, G. J. Whitman, and M. K. Markey. A model-based framework for the detection of speculated masses on mammography. *Medical physics*, 35:2110–2123, 2008.

M. Samulski, R. Hupse, C. Boetes, R. D. M. Mus, G. J. Heeten, and N. Karssemeijer. Using computer-aided detection in mammography as a decision support. *European radiology*, 20(10): 2323–2330, 2010.

D. R. Sanchez. *Breast modelling for multi-modal mammographic correspondence*. PhD thesis, Universitat de Girona, 2005.

G. Scaperrotta, C. Ferranti, C. Costa, L. Mariani, M. Marchesini, L. Suman, C. Folini, and S. Bergonzi. Role of sonoelastography in non-palpable breast lesions. *European radiology*, 18(11):2381–2389, 2008.

B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000.

F. Schwenker and E. Trentin. Pattern classification and clustering: a review of partially supervised learning approaches. *Pattern Recognition Letters*, 37:4–14, 2014.

C.-W. Seah, I. W. Tsang, and Y.-S. Ong. Transductive ordinal regression. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1074–1086, 2012.

V. Sharma and S. Singh. CFS–SMO based classification of breast density using multiple texture models. *Medical & Biological Engineering & Computing*, pages 1–9, 2014.

S.-Y. Shin, S. Lee, and I. D. Yun. Classification based micro-calcification detection using discriminative restricted boltzmann machine in digitized mammograms. In *SPIE Medical Imaging*, volume 115, page 151, 2014.

S. Singh and K. Bovis. An evaluation of contrast enhancement techniques for mammographic breast masses. *IEEE Transactions on Information Technology in Biomedicine*, 9(1):109–119, 2005.

E. Song, S. Xu, X. Xu, J. Zeng, Y. Lan, S. Zhang, and C.-C. Hung. Hybrid segmentation of mass in mammograms using template matching and dynamic programming. *Academic radiology*, 17 (11):1414–1424, 2010.

M. S. Soo, E. L. Rosen, J. Q. Xia, S. Ghate, and J. A. Baker. Computer-aided detection of amorphous calcifications. *American Journal of Roentgenology (AJR)*, 184(3):887–892, 2005.

C. Suarez-Ortega and J. M. Franco-Valiente. Using CAD systems and e-learning in radiologists training. In *15th IEEE International Conference on e-Health Networking, Applications & Services (Healthcom)*, pages 172–176, 2013.

J. Suckling, J. Parker, D. R. Dance, S. Astley, I. Hutt, C. Boggis, and J. Savage. The mammographic image analysis society digital mammogram database. In *2nd International Workshop on Digital Mammography*, 1994.

B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, and W.-B. Li. Kernel discriminant learning for ordinal regression. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(6):906–910, 2010a.

B.-Y. Sun, H.-L. Wang, W.-B. Li, H.-J. Wang, J. Li, and Z.-Q. Du. Constructing and combining orthogonal projection vectors for ordinal regression. *Neural Processing Letters*, pages 1–17, 2014.

L. Sun, L. Li, W. Xu, W. Liu, J. Zhang, and G. Shao. A novel classification scheme for breast masses based on multi-view information fusion. In *4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE)*, pages 1–4, 2010b.

Y. Sun. *Normal mammogram analysis*. PhD thesis, Purdue University, 2004.

Y. Sun, C. F. Babbs, and E. J. Delp. Normal mammogram classification based on regional analysis. In *45th Midwest Symposium on Circuits and Systems (MWSCAS)*, volume 2, pages 375–378, 2002.

Y. Sun, C. F. Babbs, and E. J. Delp. Full-field mammogram analysis based on the identification of normal regions. In *IEEE International symposium on biomedical imaging: nano to macro*, volume 2, pages 1131–1134, 2004a.

Y. Sun, C. F. Babbs, and E. J. Delp. A two-stage classifier system for normal mammogram identification. In *Computational Imaging II*, volume 112, 2004b.

J. S. Suri, S. K. Setarehdan, and S. Singh. *Advanced algorithmic approaches to medical image segmentation: state-of-the-art applications in cardiology, neurology, mammography and pathology*. Springer, 2002. ISBN 1852333898.

G. J. Szekely and M. L. Rizzo. Brownian distance covariance. *The annals of applied statistics*, 3 (4):1236–1265, 2009.

A. Tahmasbi, F. Saki, and S. B. Shokouhi. Classification of benign and malignant masses based on Zernike moments. *Computers in Biology and Medicine*, 41(8):726–735, 2011.

M. Tan, J. Pu, and B. Zheng. Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model. *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, pages 1–16, 2014a.

M. Tan, J. Pu, and B. Zheng. Reduction of false-positive recalls using a computerized mammographic image feature analysis scheme. *Physics in medicine and biology*, 59(15):4357, 2014b.

J. Tang, R. M. Rangayyan, J. Xu, I. E. Naqa, and Y. Yang. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *IEEE Transactions on Information Technology in Biomedicine*, 13(2):236–251, 2009.

X. Tang. Texture information in run-length matrices. *IEEE Transactions on Image Processing*, 7 (11):1602–1609, 1998.

Y. Tao, S.-C. B. Lo, M. T. Freedman, E. Makariou, and J. Xuan. Automatic categorization of mammographic masses using BI-RADS as a guidance. In M. L. Giger and N. Karssemeijer, editors, *Medical imaging: computer-aided diagnosis*, volume 6915, page 26, 2008.

R. F. S. Teixeira. *Automatic analysis of mammography images: classification of breast density*. MSc dissertation, Faculdade de Engenharia da Universidade do Porto, 2013.

J. S. The, K. J. Schilling, J. W. Hoffmeister, E. Friedmann, M. G. Ryan, and R. G. Holcomb. Detection of breast cancer with full-field digital mammography and computer-aided detection. *American Journal of Roentgenology (AJR)*, 192:337–340, 2009.

S. Timp and N. Karssemeijer. A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography. *Medical Physics*, 31(5):958–971, 2004.

J. Tkaczuk, I. Domingues, and J. S. Cardoso. Microcalcification detection in full field mammograms. In *16th Portuguese Conference on Pattern Recognition (RECPAD)*, 2010.

A. P. Torrent. *Simultaneous detection and segmentation for generic objects*. PhD thesis, Universitat de Girona, 2013.

M. Tortajada, A. Oliver, R. Marti, M. Vilagran, S. Ganau, L. Tortajada, M. Sentis, and J. Freixenet. Adapting breast density classification from digitized to full-field digital mammograms. *Breast Imaging*, pages 561–568, 2012.

A. Vadivel and B. Surendiran. A fuzzy rule-based approach for characterization of mammogram masses into BI-RADS shape categories. *Computers in Biology and Medicine*, 43(4):259–267, 2013.

H. Vainio and F. Bianchini. *Breast cancer screening*. Breast Cancer Research Vol 7. IARC Press, 2002. ISBN 9789283230076.

N. Vallez, G. Bueno, O. Deniz, J. Dorado, J. A. Seoane, A. Pazos, and C. Pastor. Breast density classification to reduce false positives in CADe systems. *Computer Methods and Programs in Biomedicine*, 2013.

O. Veksler. Star shape prior for graph-cut image segmentation. In D. Forsyth, P. Torr, and A. Zisserman, editors, *Computer Vision (ECCV)*, volume 5304 of *Lecture notes in computer science*, pages 454–467. Springer Berlin Heidelberg, 2008.

P. F. Velleman and D. C. Hoaglin. *Applications, basics and computing of exploratory data analysis*. Duxbury Press, 1981. ISBN 087150409X.

C. J. Vyborny, T. Doi, K. Shaughnessy, H. M. Romsdahl, A. C. Schneider, and A. A. Stein. Breast cancer: importance of spiculation in computer-aided detection. *Radiology*, 215(3):703–707, 2000.

H. Wang, J.-B. Li, L. Wu, and H. Gao. Mammography visual enhancement in CAD-based breast cancer diagnosis. *Clinical Imaging*, 37(2):273–282, 2013.

J. Wang and Y. Yang. Spatial density modelling for discriminating between benign and malignant microcalcification lesions. In *10th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 133–136, 2013.

J. Wang, H. Jing, M. N. Wernick, R. M. Nishikawa, and Y. Yang. Analysis of perceived similarity between pairs of microcalcification clusters in mammograms. *Medical Physics*, 41(5):051904, 2014.

C.-H. Wei, Y. Li, and P. J. Huang. Mammogram retrieval through machine learning within BI-RADS, standards. *Journal of biomedical informatics*, 44(4):607–614, 2011.

R. Wiemker, D. Kutra, H. Heese, and T. Buelow. Identification of corresponding lesions in multiple mammographic views using star-shaped iso-contours. In *SPIE Medical Imaging*, volume 9035. International Society for Optics and Photonics, 2014.

S.-G. Wu, J.-Y. Sun, J. Zhou, F.-Y. Li, Q. Lin, H.-X. Lin, and Z.-Y. He. The value of radiotherapy in breast cancer patients with isolated ipsilateral supraclavicular lymph node metastasis without distant metastases at diagnosis: a retrospective analysis of ChI.e, patients. *OncoTargets and therapy*, 7:281–8, 2014.

X. Xu, S. Xu, L. Jin, and S. Zhang. Using PSO to improve dynamic programming based algorithm for breast mass segmentation. In *Fifth IEEE International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, pages 485–488, 2010.

J. H. Youk, E. J. Son, J.-A. Kim, H. J. Moon, M. J. Kim, C. H. Choi, and E.-K. Kim. Scoring system based on BI-RADS lexicon to predict probability of malignancy in suspicious microcalcifications. *Annals of surgical oncology*, 19(5):1491–1498, 2012.

M. Yu, Q. Huang, R. Jin, E. Song, H. Liu, and C.-C. Hung. A novel segmentation method for convex lesions based on dynamic programming with local intra-class variance. In *27th Annual ACM Symposium on Applied Computing*, pages 39–44, 2012.

E. Zhang, F. Wang, Y. Li, and X. Bai. Automatic detection of microcalcifications using mathematical morphology and a support vector machine. *Bio-medical materials and engineering*, 24 (1):53–59, 2014a.

L. Zhang and W. Zhang. A comparison of different pattern recognition methods with entropy based feature reduction in early breast cancer classification. *European Scientific Journal*, 3: 303–312, 2014.

X. Zhang, N. Homma, S. Goto, Y. Kawasumi, T. Ishibashi, M. Abe, N. Sugita, and M. Yoshizawa. A hybrid image filtering method for computer-aided detection of microcalcification clusters in mammograms. *Journal of Medical Engineering*, 2013.

X.-S. Zhang. A new approach for clustered MCs, classification with sparse features learning and TWSVM. *The Scientific World Journal*, 2014:1–8, 2014.

Y. Zhang, Y. Lan, and H. Ren. Image enhancement and its effects on segmentation for mammographic masses. In *Fifth IEEE International Symposium on Computational Intelligence and Design (ISCID)*, volume 1, pages 423–426, 2012.

Y. Zhang, H. Zhang, H. Wei, J. Tang, and S. Zhao. Multiple-instance learning with instance selection via constructive covering algorithm. *Tsinghua science and technology*, 19(3):285–292, 2014b.

Z.-H. Zhou and M. Li. Tri-training: exploiting unlabelled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 17(11):1529–1541, 2005.

H. Zolfagharnasab, I. Domingues, and J. S. Cardoso. Breast density classification: a comparison between ordinal and traditional classification. In *2nd PhD, Students Conference in Electrical and Computer Engineering (StudECE)*, 2013.